B. SPECIFIC AIMS

Excellence and equity goals motivate much of American educational policy. We expect our schools to improve the nation's educational achievement and attainment levels, even as we ask them to narrow existing social and educational inequalities. These two goals are not always mutually reinforcing. Some policies boost average academic achievement even as they broaden educational inequalities. Others depress academic achievement even as they narrow inequalities (Hochschild & Scovronick 2003).

The twin goals of excellence and equity should lead policy-makers to be interested in both the *average* effects of educational policies and their *distributional* consequences. But although developmental science suggests that many interventions should have heterogeneous effects, most educational evaluation research focuses on the estimation of mean treatment effects either for the population at large or for particular subgroups of interest. This project aims to introduce statistical techniques related to the distributional analysis of quantitative data to educational policy research. Building on recent innovations in quantile regression, quantile treatment effect (QTE) estimation, and other explicitly distributional approaches (e.g., Abadie, Angrist, & Imbens 2002; Angrist, Chernozhukov, & Fernandez-Val 2006; Firpo 2007; Chernozhukov, Fernandez-Val, & Melly 2009; Athey & Imbens 2006), we will develop and apply new methods for modeling the effects of educational interventions on the distributions of outcomes.

By applying distributional analysis techniques to settings that are theoretically, developmentally, and methodologically distinctive, we hope to demonstrate their broad relevance to researchers who are interested in policy and youth development. The project will pursue the following key aims:

Aim 1: Estimate the distributional effects of the federally-funded Head Start program on the academic and behavioral outcomes of low-income 3 and 4 year old children. Use distributional methods to test Project I's compensatory and skill-begets-skill hypotheses regarding heterogeneous treatment impacts across the distribution of student outcomes. Compare the distributional effects uncovered by the QTE with Project I's group-based mean estimates of impact heterogeneity to better understand the merits of these two approaches.

Aim 2: Analyze the effects of a school voucher experiment for low income elementary school students in New York City on the distribution of academic skill. Test competing hypotheses that school vouchers will boost achievement either primarily at the bottom of the distribution (thus mitigating inequality) or primarily at the top of the distribution (thus exacerbating inequality). Estimate the extent to which means within subgroups reveal the distributional effects uncovered by the QTE.

Aim 3: Use distributional methods to investigate the consequences of California's policy effort to enroll more 8th graders in Algebra, paying particular attention to the differential effects of curricular policies across the distribution of student motivation and academic achievement.

Aim 4: Use QTE to measure the effects of high school exit exam requirements on the distribution of student test scores. Compare the effect of the exit exams on scores near this high-stakes threshold, where we would expect the largest effects, to other points in the distribution, where we would expect smaller or negative effects.

Aim 5: Promote the use of distributional methods among education policy researchers. Create new software applications, post web tutorials to the network website, host instructional seminars, and develop best practices for analyzing the distributional effects of human capital interventions.

Our analyses of Head Start and school vouchers, like many other studies using the QTE or other explicitly distributional approaches, use data from randomized control trials. Our analyses of 8th grade Algebra requirements and high school exit exams will apply QTE and other distributional estimators (e.g., Athey & Imbens 2006; Abadie 2005; Heckman, Ichimura & Todd 1998) to observational data on human capital interventions. Doing so creates new challenges regarding the identification of causal effects. However, by refining methods for estimating distributional effects in non-experimental settings and outlining best practices for their application, we hope to enrich the QTE's usefulness for researchers who study educational and other social policies. By demonstrating the applicability of distributional methods in a wide range of educational settings and providing researchers with the tools to undertake similar analyses, we will facilitate the use of distributional methods in other youth development research settings.

C. RESEARCH STRATEGY

C.1 Significance

In education, one size often does not fit all. Educational interventions that work to improve some students' educational outcomes may have unintended and unnoticed negative effects on other students. The theory of child/policy fit provides an important conceptual tool for understanding these uneven effects. Maintaining that "behavior, motivation, and mental health are influenced by the fit between the characteristics individuals bring to their social environments and the characteristics of these social environments", Eccles et al. (1993, p. 91) suggest that the effects of educational settings depend upon the match between these settings and the needs and abilities of students. If the needs and abilities of students vary, the theory of child/policy fit implies that the effects of a given educational intervention should also vary.

Given the central role that equity goals play in American education, it is not surprising that educational practitioners and educational policy researchers often speculate about the heterogeneous effects of educational policies across the ability and income distributions. However, until recently, analysts have lacked the methodological tools to conduct rigorous distributional analyses of educational treatment effects. Quantile regression and QTE estimators, which explicitly model differences in treatment effects across a continuous distribution, provide a powerful new framework for conducting distributional analyses. These methods have begun to make inroads elsewhere in social policy research (Heckman, Smith, & Clements 1997; Koenker & Bilias 2001; Bitler, Gelbach, & Hoynes 2006b, 2008), but, with few exceptions (e.g., Angrist, Lang, & Oreopoulous 2009) they are not yet utilized in the educational policy literature.

Recent distributional analyses of school accountability policies demonstrate the potential that these techniques offer for educational researchers. There is strong evidence suggesting that the *No Child Left Behind* act and other school accountability policies have small positive average effects on student achievement (Carnoy & Loeb 2002; Dee & Jacob 2009; Figlio & Rouse 2006; Hanushek & Raymond 2004; Wong, Cook, & Steiner 2009). But new distributional analyses indicate that these average effects tell only part of the story. Accountability policies give schools incentives to direct teacher attention and other educational resources at students near policy-mandated proficiency thresholds. By estimating policy effects on the test score distribution, Neal and Schanzenbach (2010) and Reback (2008) show that these policies have strong positive effects on students near the proficiency threshold, and weaker effects on higher- and lower-achieving students. These effects might well have been missed had these researchers instead focused on estimation of average effects for groups thought of as being low achievers, such as low SES children.

Inspired by the theory of child/policy fit, we hypothesize that unequal distributional effects of this sort are a more general phenomenon in education. In this project, we will analyze the distributional consequences of four distinct educational interventions: Head Start, school vouchers, algebra-for-all middle school curricula, and high school exit exams. The theory of child/policy fit suggests that each of these interventions has sharply heterogeneous effects at different points in the skill distribution that are obscured by simple analyses of overall mean differences or differences in means across subgroups. By undertaking these analyses, developing new distributional analysis techniques, and educating researchers in related methods, we hope to facilitate the introduction of distributional analysis more broadly in educational and youth development research.

C.1.a Effects of Head Start on the Skill and Behavioral Distributions

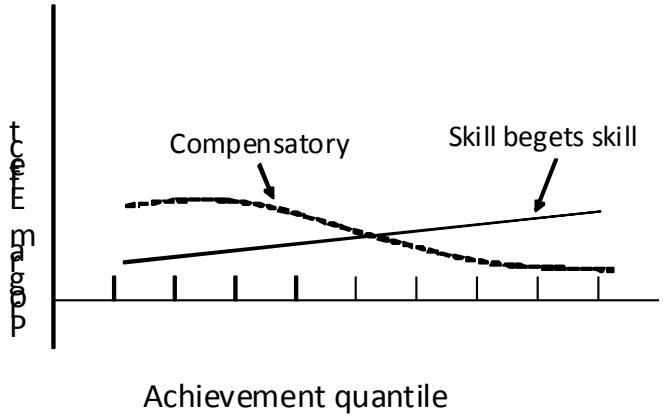
Created in 1965, the federal Head Start program is among the more prominent educational initiatives in the US. By giving matching grants to programs providing comprehensive early education, health care, and nutritional services to poor children, and parental training to their parents, Head Start aims to raise educational attainment levels and narrow educational inequalities. Head Start now enrolls more than 900,000 children and has an annual operating budget of nearly $7 billion (US DHHS, Office of Head Start 2008).

Head Start and other efforts to expand preschool education are predicated on the notion that the positive effects that educational interventions have on young children multiply across their life course (Cunha et al. 2006; Heckman 2006, 2007). For example, Knudsen, Heckman, Cameron, & Shonkoff (2006) summarize research across various fields (economics, neurobiology, and developmental psychology), concluding that early experiences are key to later development, and that early in life is the most promising period for investment in disadvantaged children. Experimental data from the Perry Preschool program, the Abecedarian Project, and the Chicago Child-Parent Center lend credence to this argument, demonstrating that early education programs can have positive effects on participants' academic achievement and attainment (Barnett 1996; Campbell & Ramey 1995; Currie 2001; Schweinhart, Barnes, & Weikart 1993).

However, the evidence regarding the long-term effects of Head Start is mixed (Currie & Thomas 1995, 2000; Garces, Thomas, & Currie 2002; Ludwig & Miller 2007; Deming 2009), and in 1998, Congress mandated a national evaluation of Head Start. The resulting Head Start Impact Study (HSIS) is a randomized control trial designed to determine "the impact of Head Start on children's school readiness and parental practices" as well as "under what circumstances Head Start achieves its greatest impact and for which children" (Puma et al. 2005) by randomly assigning children on waiting lists to be eligible for or ineligible for Head Start. We will use these experimental data to test the distributional hypotheses we lay out below.

Although existing Head Start analyses have focused primarily on average effects, population-based analyses indicate that low-achieving children stand to gain the most by enrolling in early education (Magnuson et al. 2004; NICHD 2003). This effect may be particularly pronounced in the context of Head Start, since the program's curricula is explicitly geared toward remedying the skills deficits that often disadvantage poor students at the beginning of elementary school, and towards improving the parenting practices of their parents (Puma et al. 2005). We hypothesize, therefore, that the effects of Head Start availability vary across the skill distribution, with students at the bottom of the distribution experiencing the most pronounced positive academic and language effects. Similarly, we expect the program to improve the behavior of children who are relatively prone to behavior problems. This prediction, which parallels the "compensatory hypothesis" outlined in Project I, is shown by the dotted line in Figure B.1, which plots the predicted effect of Head Start (Y-axis) across the achievement distribution (X-axis).

Figure B.1 Two hypotheses regarding the distributional effect of Head Start



In contrast, the "skills-begets-skills" hypothesis outlined in Project I suggests that the positive effects of Head Start may be concentrated at the top of the skill distribution. Research on learning trajectories in elementary school and beyond indicates that since academic skills are cumulative, achievement inequalities tend to widen as children progress through school (Stanovich 1986). This theory may also be relevant to early education. For example, it is plausible that students who have basic language and numeric competencies are more able to participate fully in the Head Start curriculum than their academically or developmentally-delayed peers. In this case, the distributional effects of Head Start may resemble the "skills-begets-skills" hypothesis outlined as the solid line in Figure B.1. We believe that the overlap between our HSIS analyses and those proposed by Project I represent an opportunity for collaboration. By writing a paper comparing the results of our analyses of the distributional effects of Head Start with the results emerging from Project I's estimates of the average effects on particular subgroups, we will gain new insight into the strengths and limitations of distributional analysis in educational research as compared to the standard mean impacts approach.

C.1.b Competing Hypotheses Regarding the Effects of School Voucher Programs

Arguing that traditional public schools are monopolistic and inefficient, school voucher proponents aim to create more vibrant educational marketplaces. By broadening the educational choices available to parents and students and creating incentives for schools to improve, voucher programs promise to boost educational outcomes for students who might otherwise have no choice but to enroll in low-quality public schools (Chubb & Moe 1990; Friedman & Friedman 1980).

School reformers have launched a handful of voucher programs across the United States over the past two decades in an attempt to demonstrate the effectiveness of this approach. In 1997, the School Choice Scholarships Foundation initiated one such program in New York City, offering three-year scholarships worth $1,400 a year to a randomly selected group of low income children in grades K–4. This program's random assignment design makes it possible to distinguish the effects of voucher availability from the potentially confounding characteristics of families who self-select into voucher programs. More than 10,000 eligible students applied; around 2,600 scholarship recipients were chosen via lottery. In the years that followed, researchers from Mathematica Policy Research (MPR) and the Harvard University Program on Education

Policy gathered enrollment and achievement data from students in the treatment and control groups.

Analyses of the New York City voucher experimental data clearly indicate that vouchers influence school choice. Students randomly selected to receive a voucher were several times more likely than their peers in the control group to attend private schools (Mayer et al. 2002). It is much less clear, however, whether the voucher offer had an effect on student achievement. Several studies find that the voucher offer had a small positive effect on the academic achievement of African-American recipients (Barnard et al. 2003; Howell et al. 2002; Peterson & Howell 2004). However, for all other racial and ethnic groups the voucher program had no effect. Furthermore, subsequent analyses suggest that the observed effects for African Americans are sensitive to the definition of racial and ethnic categories and hold only when controlling for students' initial characteristics (Krueger & Zhu 2004a, 2004b). We suspect that the weak average effects of vouchers disguise larger (and possibly contradictory) voucher program effects for high or low achieving students. The theoretical literature surrounding school choice suggests two competing hypotheses regarding the effects of voucher programs on student achievement.

The "common school hypothesis" holds that voucher programs benefit low-performing students disproportionately. This hypothesis is grounded in the literature on the effects of Catholic schools. Since Catholic schools are typically smaller than public schools and their curricula are often relatively undifferentiated, low-performing students tend to benefit disproportionately from enrolling in Catholic schools (Hoffer, Greeley & Coleman 1985; Coleman, Hoffer, & Kilgore 1982; Evans & Schwab 1995; Morgan 2001). By providing a mechanism for students to opt out of neighborhood public schools and into Catholic and other private schools, voucher experiments attempt to make the positive achievement effects associated with Catholic schools more broadly available. Assuming that the Catholic school effects uncovered in national data analyses generalize to the schools that voucher recipients chose, the "common school hypothesis" suggests that voucher school programs have positive effects on students at the bottom end of the academic achievement distribution.

However, the "stratifying hypothesis" suggests that voucher programs magnify educational inequalities. Voucher program advocates take it for granted that parents use school choice to maximize their children's educational success. In practice, however, many parents make school choice decisions based on the convenience of the school's location, its disciplinary style, and its religious affiliation (Elacqua, Schneider, & Buckley 2006; Weiher & Tedin 2002; Buckley & Schneider 2002; Hastings, Kane, & Staiger 2005). Hastings, Kane, & Staiger hypothesize the effects of voucher programs are contingent on the quality of the school choices that families make. For students whose families make school choices on the basis of academic quality, voucher programs may have positive effects. But for students whose families make school choices based on other factors, vouchers may have zero or negative effects. If these educational preferences vary with student academic achievement, voucher programs may provide a boost for students who are high achievers pre-voucher, even as they negatively affect test scores for low achievers.

These competing hypotheses have not been thoroughly investigated in an explicitly distributional fashion. Using QTE methods to analyze this experimental data will add to our understanding of what parts of the distribution benefit from access to vouchers. We can test systematically whether the variation in effects for various subgroups found in previous literature and reported in the Mathematica report captures the important variation across the distribution of test scores.

C.1.c Middle School Curricular Change and Student Achievement and Motivation

Driven by complaints about the rigor of American educational standards (National Commission on Excellence in Education 1983, National Diploma Project 2004, National Governor's Association 2005), as well as concerns about academic tracking (Lucas 1999; Oakes 1985; Powell, Farrar, & Cohen 1985), a national movement is underway to make middle and high school curricula more rigorous. Most US states now require students to complete college preparatory curricula in order to earn high school diplomas (Achieve, Inc. 2007), and transcript studies reveal a pronounced trend toward advanced course-taking (Dalton et al. 2007).

These shifts are particularly pronounced in California, where the State Board of Education has moved to require all public school students to take Algebra I by 8th grade. While this proposal is currently pending appeal in state courts (Rosin, Barondess, & Leichty 2009), California schools are already moving to meet the mandate. Between 1999 and 2008, the proportion of California 8th graders enrolled in algebra more than tripled from 16 percent to 51 percent.

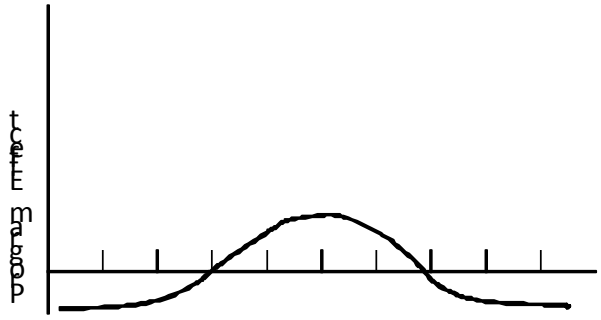
We propose to evaluate the distributional consequences of this curricular change on achievement and motivation using rich, longitudinal student-level data from three high-poverty ethnically-diverse Southern

California school districts. One of the districts that we propose to study has been at the leading edge of California's 8th grade algebra movement, and today it places nearly all 8th graders into algebra courses. While 8th grade algebra enrollment rates have also risen in the other two districts, the increase came later and has been more gradual. As a result, these districts provide a natural experimental setting for analyzing the effects of 8th grade algebra course placement on 8th graders' mathematics achievement and motivation, as well as their later academic trajectories.

Research strongly indicates that challenging course placements have positive average effects on student learning and educational attainment (Attewell & Domina 2008; Chaney, Burgdorf & Atash 1997; Gamoran & Hannigan 2000). However, curricular reforms may have unintended effects (Allensworth et al. 2008) and it remains unclear whether all students benefit equally from enrolling in more advanced courses. Large-scale changes in course-taking patterns typically require changes in the design and delivery of course materials, as classes that were once taught to groups of students whose skill levels were relatively homogenous are now delivered to heterogeneous groups. Faced with students who are less uniformly well prepared for their courses, teachers in newly required courses may cover less material than in the past. Even with these changes, newly required courses may still be too advanced for some students (Loveless 1999; Rosenbaum 1999).

Our unique data enable us to look at distributional effects of curricular reform on academic achievement and motivation. We will use matching and nonlinear differences in differences methods to assess the effects of the increase in algebra enrollment on student test scores and achievements. We hypothesize that California's effort to enroll all 8th graders in Algebra has positive average effects on student achievement and motivation, since it places greater cognitive demands on students at a crucial point in their development (Eccles et al. 1993; Giedd 2008). However, we suspect that the reform may have negative effects at the top and bottom of the skill and motivation distributions. If the push for curricular intensification leads to increased heterogeneity in 8th grade Algebra classrooms, it may have a negative effect on high-achieving students who would have already enrolled in Algebra without the policy. At the opposite end of the skill distribution, approximately one-third of California 8th graders who enroll in algebra fail to demonstrate proficiency on end-of-the year algebra courses (Rosin, Barondess, & Leichty 2009). The 8th grade algebra requirement may discourage these students and have negative effects on their math achievement. Figure B.2 illustrates this hypothesis for achievement. In addition, we will estimate the distributional effects of curricular change previously unexamined measures such as student feelings of self-efficacy and several distinct intrinsic and extrinsic motivation measures (including mastery, performance approach and performance avoidance goal orientations).

Figure B.2 Hypothesized distributional effect of 8th grade algebra



C.1.d Existing Evidence about High School Exit Exams and Gaps in Knowledge

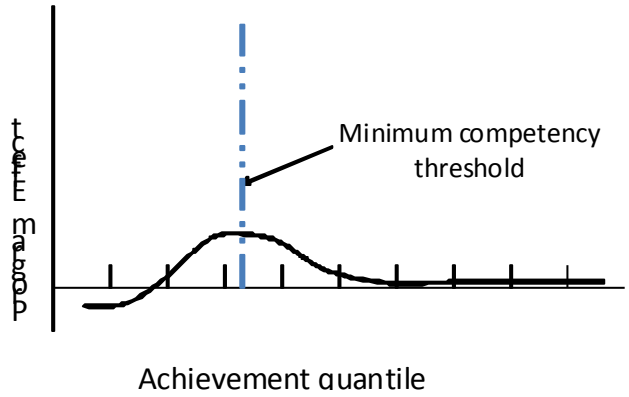
Until recently, high school students in the US earned diplomas exclusively by accumulating course credits. Over the last three decades, however, a growing number of states implemented high school exit exams and mandated that students pass these tests in order to earn diplomas. Today, high school exit exams are in place in 22 US states, and approximately two-thirds of US high school students must pass an exit exam to graduate from high school. California's high school exit exam (CAHSEE) was mandated by the state legislature in 1999. While all students in the high school class of 2003 took the exam, they could graduate without passing. For students in all subsequent classes, passing CAHSEE has been a requirement for high school graduation.

Several studies indicate that high school exit exams have negative effects on students' odds of earning high school diplomas (Dee & Jacob 2007; Martorell 2004; Papay, Murnane & Willet 2008; Warren & Edwards 2005). In California, nearly one-third of students fail to pass the exit exam on their first attempt (Human Resources Research Organization 2008) and regression discontinuity analyses reveal that students who barely fail the exam on their first attempt are considerably more likely to drop out of high school than they would have been had they barely passed (Reardon et al. 2009). On average, Warren, Jenkins, & Kulik (2006) estimate that when states implement high school exit exams, high school graduation rates fall by approximately 2 percentage points. We will analyze the effects of CAHSEE on student graduation odds using official administrative

enrollment and graduation data for a large southern California district. Our analyses will pay particular attention to the exam's distributional consequences, estimating policy effects across the distribution of students' latent probability to earn a high school diploma.

In addition, our work will supplement the uneven existing literature on the effects of exit exams on student achievement. By establishing a minimum competency threshold and giving students strong incentives to clear it, high school exit exams aim to raise students' academic achievement. In the proposed project, we will investigate the effects of California's exit exam policies across the academic achievement distribution. While some evidence points to a positive relationship between the existence of high school exit exams and student achievement (Bishop 1997; Bishop, Moriarty, & Mane 2001), the most rigorous available studies indicate that exit exams have no causal effect on achievement (Jacob 2001; Reardon et al. 2009). By estimating the average effects of high school exit exams—or, in the case of the Reardon et al. study, estimating the local effects of high school exit exams at the passing threshold—these studies neglect the possibility that high school exit exams have different effects on students with differing levels of academic skill. In the existing literature, only Grodsky, Warren, & Kalgorides (2009) investigate the effects of exit exams across the skill distribution, using nationally representative NAEP data; these national comparisons may obscure distributional effects if the minimum competency thresholds vary across states.

Figure B.3 Hypothesized distributional effect of CAHSEE



Our hypothesis, illustrated in Figure B.3, is that exit exams have varying effects on the academic skill distribution. By raising the stakes associated with minimum competency on the skills it assesses, CAHSEE may encourage students who expect to be close to the minimum competency threshold to redouble their academic efforts. On the other hand, we do not expect CAHSEE to have a positive effect on the academic achievement of the large number of students who expect to score well above the minimum competency threshold. In fact, the exit exam may even have a modest negative effect on some of these relatively high-achieving students' academic subsequent achievement, if they are motivated by the desire to earn a high school diploma. Lastly, we anticipate negative effects for students who expect to score well below the threshold. Existing literature on high school exit exams and other accountability interventions fail to examine effects on the full distribution. Our work comparing effects across the skill distribution will fill an important gap in this literature.

C.2 Innovation

By explicitly modeling the unequal effects of Head Start, school vouchers, middle school curricular requirements, and high school exit exams on academic, behavioral, and socioemotional outcomes, our distributional analyses will substantially enrich our understanding of each of these four interventions. These studies will potentially uncover effects that would be obscured by standard mean effects analysis, given hypotheses about signs and magnitudes of effects varying for different parts of the skill distribution. Further, given concerns about equity behind interventions such as Head Start and school choice programs, evidence about the distribution of these effects would be policy relevant. Evidence of disproportionate large positive Head Start effects on the bottom of the student performance distribution would strengthen the case that investments in early education are particularly effective mechanisms for narrowing academic achievement gaps. Our analyses of the distributional consequences of school vouchers will speak to long-standing debates about the relationship between school choice and inequality. Investigating the ways 8th grade algebra assignments affect the distributions of student math skills and student dispositions toward school will provide a much clearer picture of what part of the distribution benefits or loses when curricular standards rise. Finally, distributional analyses of the effects of exit exams on student achievement will shed light on the ways high stakes educational incentives affect academically marginal high school students.

A common practice in program evaluation is to look for treatment effect heterogeneity by estimating mean differences between control and treatment groups (perhaps with regression adjustments). We will also explore the extent to which mean differences by subgroup capture the heterogeneity in the effects of these interventions across the distribution of skill. We will benefit particularly from findings in Project I in our choice

of subgroups, and our results will inform the interpretation of Project I's findings. We will also have many chances to encourage the rest of the INID network to use these methods in future work where appropriate, and aid in the ultimate goal of identifying promising interventions.

Specific Aim 5 of this project has the broader goal of introducing distributional analysis methods to researchers in educational policy beyond the Irvine Network on Interventions in Development. QTE and other distributional estimation methods have begun to exert an influence in applied microeconomics and some areas of social policy research. However, even though these methods are highly complementary with central developmental notion of child/policy fit, they are underutilized in research related to educational policy and youth development. By applying distributional analysis to a wide range of educational interventions, we hope to demonstrate the broad applicability of distributional methods to educational policy research and encourage other educational researchers to use these techniques. We will accomplish this by presenting our work in a wide range of venues, as well as by making computer code and web tutorials available on the network website. Without transparent computer code and examples of their use, new methods can languish. For example, while Athey and Imbens (2006) was published in a top economics journal more than 4 years ago, these techniques remain underutilized in economics and largely unknown in other disciplines.

We will also facilitate the broader adoption of distributional analysis by applying new techniques for estimating distributional effect sizes using observational data. In our analyses of 8th grade Algebra placements and high school exit exams in California, identification of causal effects will be far more troublesome than it will be in our analyses of randomized Head Start and school voucher experiments. However, randomized control trials are still relatively rare in education, and many important educational interventions are not amenable to experimental analysis. As a result, quasi-experimental methods will likely remain prominent in educational policy research. We anticipate, therefore, that applying QTE estimation and other related distributional approaches to observational data will be a particularly important component in our effort to bring these techniques to educational research.

C.3 Approach

In this section, we first describe the distributional methods that we will use. We then discuss the approach for each specific aim, and present our preliminary findings.

The potential outcomes model (e.g., Rubin, 1974; Holland, 1986) provides a framework for estimation of the effects of a treatment. Each individual *i* has two potential outcomes, Y_{1i} and Y_{0i} (for our purposes, a test score or index of student behavior or motivation). Person *i* has outcome Y_{1i} if assigned to the treatment group and outcome Y_{0i} if assigned to the control group. D(i) denotes the group that *i* is assigned to in a randomized experiment. If person *i* is assigned to the treatment group, then D(i) = 1, and if person *i* is assigned to the control group, D(i) = 0; the treatment effect on person *i* is defined as $\delta_i = Y_{1i} - Y_{0i}$. The fundamental evaluation problem is that the same person cannot simultaneously be in the treatment group and the control group.

C.3.a Quantiles, Average Treatment Effects, and Quantile Treatment Effects

Let Y be a random variable with a cumulative distribution function (CDF) F(y), where $F(y) = \Pr[Y \leq y]$. Then, the *q*th quantile of the distribution F(y) is defined as the smallest value y_q such that $F(y_q)$ is at least as large as *q* (e.g., $y_{0.5}$ is the median). Now consider two (marginal) distributions F_1 (the CDF for the potential outcomes if D = 1), and F_0 (the CDF for the potential outcomes if D = 0). We define the difference between the *q*th quantiles of these two distributions as $y_q = y_{q1} - y_{q0}$, where y_{qd} is the *q*th quantile of distribution F_d.

The joint distribution of (Y_{0i}, Y_{1i}) is not identified without assumptions. However, if program assignment is independent of the potential outcomes, the difference in means, or average treatment effect, $\delta = E[\delta_i] = E[Y_1] - E[Y_0]$, is identified because each expectation requires only one of the two marginal distributions. Similarly, identification of the marginal distributions implies identification of the quantiles y_{qd}, and thus identification of the differences in their quantiles, $y_q = y_{q1} - y_{q0}$. In fact, in the experimental settings of Specific Aims 1 and 2, the quantile treatment effect (QTE) is the estimator of this difference in the quantiles of the two marginal distributions. For example, in these experimental studies, we consistently estimate the QTE at the 0.50 quantile by subtracting the control group's sample median from the treatment group's sample median. Graphically, QTE estimates are the differences in the inverse CDFs of the outcome for the treatment and control groups. For an example, see section *C.3.f* below and QTE for the effects of being assigned to get an offer of a private school voucher (Figure C.1).

C.3.b When are QTE Equal to the Distribution of Individual Treatment Effects (DOTE), How to Test for This, and How to Bound the DOTE?

It is important to note that the QTE at quantile *q* need not equal the treatment effect for an individual located at quantile *q* of the control group. Thus, while our hypotheses are about effects for individuals of various cognitive abilities, our QTE results will represent effects on the whole distribution. Only with further assumptions, which are perhaps undesirable, can we conclude that the QTE is the treatment effect for a particular individual. One assumption that allows for treatment heterogeneity is rank preservation (e.g., Heckman, Smith, & Clements 1997), under which a person's location in the distribution is unchanged by the treatment. In this case, the QTE are the same as the distribution of individual treatment effects.

We will test the rank preservation assumption with a test developed in the working paper version of Bitler, Gelbach, & Hoynes (2008) using the experimental data from Specific Aims 1 and 2. This test looks at a necessary but not sufficient condition for rank preservation, that the distribution of observable characteristics across the treatment and control group outcome distributions be the same up to sampling variation within various ranges. Knowledge of evidence against rank preservation in experimental data is important. Evidence against rank preservation in experimental data, where concerns about sample selection should not be important, suggests it is even less likely to hold in observational data. Further, some leading instrumental variable quantile regression estimators devised to deal with heterogeneity in the presence of endogeneity (e.g., Chernozhukov & Hansen 2005, 2006) rely on rank preservation or rank similarity. An alternative approach is to use the QTE to bound the distribution of treatment effects. Fan & Park (forthcoming) have devised a method for putting informative bounds on the distribution of treatment effects; we will estimate these bounds in conjunction with our QTE using the experimental data from Specific Aims 1 and 2.

Furthermore, we note that even if rank preservation does not hold, QTE suffice for many interesting comparisons and provide a rich understanding of the distributional effects of human capital interventions. For example, Abadie, Angrist & Imbens (2002) point out that any analysis aimed at thinking about whether an intervention leads to more or less inequality requires only the QTE. Additionally, QTE provide an opportunity to assess the ability of theoretical model predictions to match empirical estimates of the effects of various programs (e.g., Bitler, Gelbach, & Hoynes, 2006b).

C.3.c Testing whether Mean Effects in Subgroups Explain QTEs on the Full Distributions

A common strategy in program evaluation for examining heterogeneous effects is to compute mean treatment effects separately for specific subgroups. Policy makers value this if subgroup means reveal heterogeneity and can thus potentially be used to define groups that benefit from the interventions (or for targeting). Researchers value subgroup means as a way to link theoretical predictions about who should gain or lose from an intervention, or alternatively, in observational studies as a way to focus attention on groups where the probability of being affected by the intervention is large.

We will use the methods from Bitler, Gelbach, & Hoynes (2010) to test whether mean effects within such subgroups (or others that theory suggests are good predictors of being someone who gains or loses from the interventions) can explain the heterogeneity uncovered by QTE. We first will construct QTE within subgroups of interest (e.g., based on SES or previous test scores as suggested by some of the hypotheses laid out in sections *C.1.a-d* above). We then examine whether members of these subgroups are concentrated in various points of the outcome distribution for the treatment and control groups. For example, this can be done by seeing whether the SES composition within each percentile of the treatment or control distribution varies across the distribution. We will also construct synthetic treatment group distributions by adding the mean treatment effect within subgroup to the control group observations, and calculate inverse CDFs. We will compare these synthetic inverse CDFs to the true treatment group CDFs, using a statistic suggested by Chernozhukov & Fernandez Val (2005). Intuitively, the more similar the synthetic and true treatment group distributions, the more closely these mean treatment effects within subgroup come to explaining the heterogeneity in the full unconditional distribution. This test will be particularly useful for understanding our findings regarding the distributional effects of Head Start in comparison to the findings that emerge from Project I. If we find little variation within subgroups across the distribution and/or that the subgroup members are concentrated in various parts of the outcome distribution, we will conclude that traditional mean subgroup analyses are sufficient to estimate the heterogeneity in the effects of Head Start. On the other hand, if we find more variation within than across subgroups or if we find subgroups are not systematically concentrated at various points in the distribution, we will conclude that these traditional subgroup mean impacts are missing important effects of Head Start on the distribution of achievement.

C.3.d Distributional Analyses Using Observational Data

The preceding discussion focuses on the estimation of distributional effects in experimental settings, in which random assignment insures that program assignment is independent of the potential outcomes. The randomized control trials we propose to analyze in Specific Aims 1 and 2 fit this description. However, for many questions of interest to educational and youth development researchers, such experimental analyses are not practical. Therefore, in Specific Aims 3 and 4 we propose to apply similar distributional analysis techniques to observational data.

In a quasi-experimental setting, if treatment assignment is as good as random, we can also estimate QTE simply by taking the difference between treatment and control groups at each point in the distribution. For example, our analysis of the effects of the CAHSEE exam compare the first cohort of students who were subject to this high stakes exam with the preceding cohort, which took the exam as a low-stakes test. In this case, the estimated QTE is the differences in the inverse CDFs for the treatment and control groups. Thus, for example, the 0.50 quantile treatment effect is given by the vertical difference in the inverse CDFs at the median.

Difference in difference methods are another common approach to policy evaluation in the absence of experimental data. Under the assumption that average outcomes for the treatment and control group would have followed parallel paths over time, differences in differences models uncover consistent estimates of the true underlying average treatment effect. However, this parallel trends assumption is undesirable if pre-treatment characteristics associated with dynamics are unbalanced among groups (e.g., Ashenfelter's dip). Abadie (2005) extends the standard difference in differences estimator to one where even in the face of such differences, a two-step strategy leads to consistent estimates of the average treatment effects which vary with changes in observed characteristics. This approach allows calculation of average treatment effects for different values of the covariates. Thus, one could estimate the average treatment effect given a specific value of one of the *X*s, rather than an average treatment effect given a specific value of the propensity score.

We will also explore extensions of propensity score methods (e.g., Hahn, Todd, & van der Klaaw 2001; van der Klaaw 2002) to differences in differences settings (e.g., Heckman, Ichimura & Todd 1998; Heckman, Ichimura, Smith, & Todd 1998). Such methods first estimate the probability that each observation is treated conditional on the *X*s (i.e., the propensity score), and either use its inverse to weight comparisons match on the propensity score to get the counterfactual for treatment group members. Propensity score methods also ensure that only "comparable" control group members are used as a comparison for each treatment group member. Blundell et al. (2004) extend propensity score matching to differences in differences models. They estimate two propensity scores, one for being in the treatment group, and the second for being in the pre- versus post-period. The combination of these propensity scores balances the four groups (treatment group pre-program, treatment group post-program, control group pre-program, and control group post-program).

We will also apply the "changes in changes" model of Athey and Imbens (2006) for estimating nonlinear differences in differences models. This model, which is less restrictive than standard differences in differences model, yields a simple prediction for the counter-factual distribution of the treatment group in the absence of the treatment, $F_{t2} = F_{t1}(F^{-1}_{c1}(F_{c2}(y)))$. Then the estimated change in change treatment effect is given by the difference between this counterfactual distribution and the actual treatment group distribution at a given quantile. Figure C.2 below shows the results of a preliminary changes in changes analysis.

C.3.e Approach for Specific Aim 1 and Progress to Date: Head Start Impact Study

The data from the Head Start Impact Study (HSIS) are not yet publicly available, although the final report was released in January and the public use data are expected soon. These data follow for 3 years the outcomes of a nationally representative sample of 3 or 4 year old Head Start applicant children who were experimentally assigned to a treatment (eligible for Head Start) or control group (not eligible).

After checking that the treatment and control groups are balanced across the distributions, we will test the distributional hypotheses laid out above in section *C.1.a* by estimating QTE on year 1 and 2 additional years of follow-up data from the Head Start Impact Study. We will look at distributional effects of Head Start on a subset of the many academic and social-emotional measures collected in the HSIS. For early language, we will look at the Peabody Picture Vocabulary Test (adapted), and for early literacy the Woodcock Johnson III Letter Word Identification. For early mathematics, we will examine the Woodcock Johnson III Applied Problems. For learning-related behavior, we will consider the Leiter-R Sustained Attention Task (adapted). We will also examine the effects of assignment to Head Start on parenting behavior and health outcomes, also using QTE. Dichotomous measures will be analyzed using Kordas's (2006) method. We will test the skills beget skills

hypothesis by seeing if the effects of being assigned to Head Start benefit the top of the distributions of academic, behavioral, or parenting distributions among children eligible for Head Start. We will test the compensatory hypothesis that instead Head Start will benefit the low end of the distribution by seeing if the QTE at the lower end of the distribution are positive.

As laid out in further detail in Project I, developmental theory suggests differences in the effects of Head Start on achievement and social-emotional outcomes by SES, age, baseline child temperament, and race/ethnicity. Software is available from Bitler and Hoynes's earlier work to systematically test whether the overall heterogeneity can be replicated by subgroup means. We will explore the extent to which any differences across subgroups of policy interest such as children of different SES groups explain any effects we detect on the distribution. We can test to what extent the observable characteristics that developmental theory suggests are important to child/policy fit (such as those articulated in Project I) account for the heterogeneity of effects across the distribution. These tests provide insight into the ability of subgroup analyses to describe the sources of effect heterogeneity in the overall population. We can also examine the extent to which subgroup mean differences explain or obscure within group variation by estimating QTE within subgroups. Finally, we will use the Fan and Park (forthcoming) method to bound the distribution of individual treatment effects and the test from Bitler, Gelbach, and Hoynes (2006a) to test for evidence against rank preservation. Table C-1 has further details about this and the other datasets.

Table C-1: Data sets, measures, and subgroups predicted to have heterogeneous effects

Specific Aim 1	
Data set/Sample size/Experiment:	National Head Start Impact Study/4,667/Experiment
Treatment group:	Low income 3 and 4 year olds assigned to get into HS
Control group:	Low income 3 and 4 year olds assigned to not get into HS
	Direct child assessments
Early language:	Peabody Picture Vocabulary Test (adapted)
Early literacy:	Woodcock-Johnson III Letter-Word Identification
Early mathematics:	Woodcock-Johnson III Applied Problems
Learning-related behavior:	Leiter-R Sustained Attention Task (adapted)
	Parent reports
Social-emotional:	Total Child Behavior Problems Scale, Aggressive Behavior Subscale, Hyperactive Behavior Subscale
Parenting:	Safety Devices Scale
Health insurance/health	Child has health insurance, Child obtained dental care
	Predictions about treatment effects heterogeneity by:
Subgroups:	Race/ethnicity (White/other), SES (very low/low), Age at baseline(3/4), Gender (boy/girl), Baseline temperament (difficult/not difficult)
Specific Aim 2	
Data set/Sample size/Experiment:	School Choice Scholarships Program Evaluation/2080/Experiment
Treatment group:	Low income children offered a voucher
Control group:	Low income children not offered a voucher
	Direct child assessments
Achievement measures:	Iowa Test of Basic Skills Math Composite Score
	Iowa Test of Basic Skills Reading Composite Score
	Parent reports
School characteristics:	Class size, Attended private school, Parent's satisfaction with school
Child behavior:	Suspended in last year
	Predictions about treatment effects heterogeneity by:
Subgroups:	Race/ethnicity (Black/Hispanic/other), SES (very low/low), Grade at baseline(K–4), Gender (boy/girl)
Specific Aim 3	
Data set/Sample size/Experiment:	California Motivation Project-Math (CAMP-Math)/13,829/Non-experimental
Treatment group:	Students induced to take Algebra 1 in 8th grade by school policy change
Control group:	Students not induced to take Algebra 1 in 8th grade by school policy change
	Direct child assessments
Achievement measures:	California Standards Test (CST) Mathematics, CST English Language Arts
Motivation measures:	Academic efficacy, Performance avoidance goals, Mastery goals, Performance approach goals
	Predictions about treatment effects heterogeneity by:
Education	Stayed in school until grade 12, Graduated, Attended further school
Subgroups:	Race/ethnicity (Vietnamese/Hispanic/other), SES (very low/low), Gender (boy/girl), Language fluency (English proficient/not)
Specific Aim 4	
Data set/Sample size/Experiment:	Administrative data from several school districts/13,583/Quasi-experiment
Treatment group:	10th grade students who had to pass CAHSEE to graduate (2004–05 cohort)
Control group:	10th grade students who could fail CAHSEE and graduate (2003 cohort)
	Direct child assessment
Achievement measures:	California High School Exit Examination (CAHSEE), CST Mathematics, CST English Language Arts, District tests in Mathematics, English Language Arts, other subjects
	Predictions about treatment effects heterogeneity by:
Education	Stayed in school until grade 12, Graduated, Attended further school
Subgroups:	Race/ethnicity (white/Hispanic/Vietnamese/other), SES (low/high), Gender (boy/girl), Language fluency (English proficient/not)

C.3.f Approach for Specific Aim 2 and Progress to Date: Effects of Voucher

The New York City School Choice Scholarship Program (NYCSCSP) was a three year private school choice randomized experiment. As noted above, low income students (students qualified for free school lunch) in grades K–4 were eligible to apply for vouchers of $1,400 to be used towards private school tuition. While not representative of all students, these students are more similar to those in districts facing challenges to improve. We have obtained access to the NYCSCSP data (see Table C-1) we will use in Specific Aim 2 to evaluate distributional effects of a voucher experiment in New York City from Mathematica Policy Research. After checking the treatment and control groups are balanced, we will estimate QTE for the Iowa Test of Basic Skills Math and Reading composite scores and for two school characteristics: class size and an index of parental satisfaction with the child's school. We will explore the extent to which the overall heterogeneity can be replicated by subgroup means, paying attention to the characteristics suggested by the evaluators to define groups who had different gains from assignment to the treatment group as well as suggestions of theory.

Figure C.1 QTE estimates of NYCSCSP effects on first-year math test scores

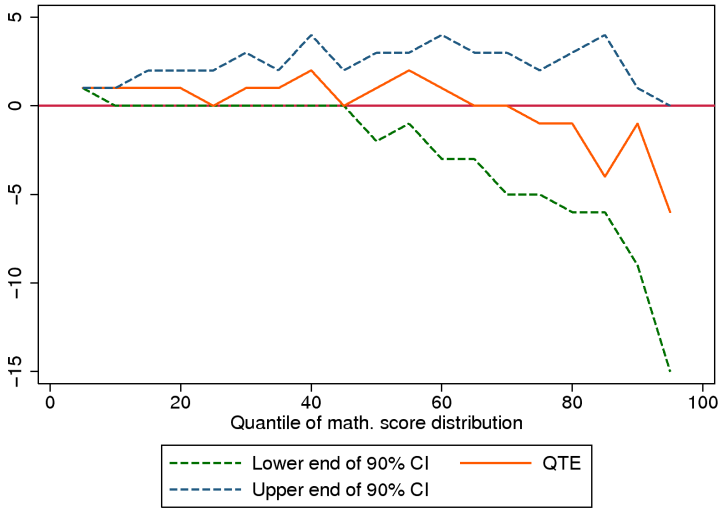


Figure C.1 shows our preliminary findings for the effects of voucher receipt on student test scores measured one year after the voucher lottery occurred. The NYCSCSP had no mean effect on student math achievement (Howell et al. 2002; Krueger & Zhu 2004a, 2004b). However, our preliminary analysis suggests that this null effect obscures a more nuanced distributional effect of school vouchers. Voucher receipt seems to have a positive effect for low-performing students, and a negative effect for high-performing students.

C.3.g Approach for Specific Aim 3 and Progress to Date: Effects of Increased Algebra Enrollment

In specific aim 3 we will look at the ramifications of California's effort to encourage all middle school students to attempt higher level math. We will examine the effects of increasing 8th grade algebra enrollment in several school districts over 2004–2009 on the distributions of test scores in math (measured by the CST, and in some cases, district tests), as well as the distribution of measures of motivation. We will use California Motivation Project – Mathematics (CAMP-Math). The current CAMP-Math data were collected between 2004 and 2006, and includes information about grades 6–12 for all students in about 500 classrooms within 14 schools with diverse ethnic and linguistic groups in districts in Southern California.[1] This dataset includes repeated measures of student academic achievement as well as data on children's and teachers' motivation collected at four points during a two-year study period. In addition, we have access to detailed data on course titles. Based on ongoing discussions with administrators in the study districts, we are confident that we will be able to use these codes to account for changes in algebra and other math course curricula over the study period. We have combined the data on motivation collected as part of CAMP with our data on test scores for students in one school district (and plan to do so for the rest of the districts and schools). We have calculated trends in underlying enrollment in algebra in this district from publicly available data from the California Department of Education. Table C-2 shows average rates of school enrollment in Algebra 1 for 8th graders for these schools and the state overall, from state-wide administrative data. This clearly shows that for our schools, there was a large increase in the share of 8th graders enrolled in Algebra 1 between 2005 and 2006.

Table C-2: 8th grade Algebra 1 enrollment, statewide and sample

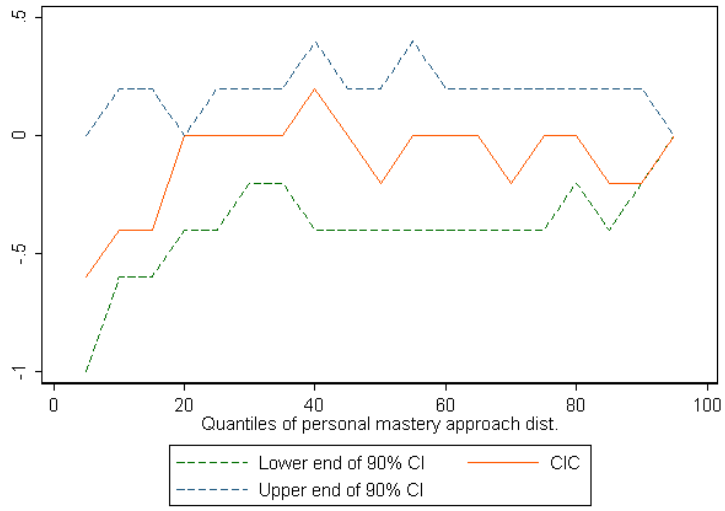
Year	2002	2003	2004	2005	2006	2007	2008
State	.33	.36	.42	.49	.52	.54	.56
District 1:							
School A	.22	.23	.24	.29	.51	.94	.94
School B	.17	.25	.23	.31	.40	.66	.88
School C	.14	.14	.28	.39	.46	.51	1
District 2:							
School A	.31	.32	.17	.25	.35	.25	.18
School B	.11	.13	.21	.22	.30	.34	.34
District 3:							
School A	.16	.27	.15	.17	.24	.27	.34

We plan to explore the effects of enrollment increases on both student achievement (CST scores) and measures of self-efficacy and achievement goals, using the various generalizations of the differences in differences model laid out above in section *C.3.d*; semiparametric and propensity score differences in differences models, and the changes in changes model. We will examine not only mastery goals, which are linked to developing competence and are thought to be adaptive, but also performance goals, which are linked to showing competence and are viewed as less beneficial. Figure C.2 shows the changes in changes estimate of the effects of being in a school which substantially increased enrollment in Algebra 1 on one of our measures of motivation: personal performance approach goals (the changes in changes methodology is discussed above in section *C.3.d*, and is a distributional analog to difference in difference methods). As Figure C.2 suggests, the increased emphasis on algebra enrollment led to few significant changes in student personal performance goals, although there is a suggestion of a negative effect in the bottom third of the distribution. Further work

[1] We will also explore differences between the CAMP schools and others in the district as well as explore how different the districts are from the rest of the state. These schools face similar challenges to others with large minority populations.

will involve using the other differences in differences estimators on various test scores and other motivation measures. Further, given the strong relationship between students' motivation and mathematics achievement (Schoenfeld 1992; Fennema 1989), we plan to examine the joint distribution of mathematics achievement and motivation.

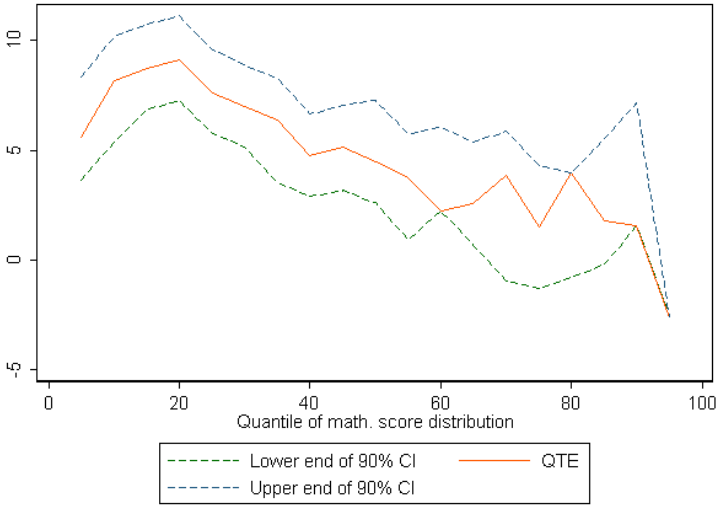
Figure C.2 Changes in changes estimate of effect of Algebra enrollment on personal mastery approach goals



C.3.h Approach for Specific Aim 4 and Progress to Date: Effects of Exit Exams

In the words of California state superintendent of instruction Jack O'Connell, the CAHSEE competency exam "measures absolutely the least our students must know as they move on to their next step in learning and earning." The exam consists of two parts: (1) a multiple choice mathematics portion, aligned with the California state math standards for 6th and 7th graders, as well as Algebra I, and (2) an English language arts (ELA) portion, which uses multiple choice questions and one essay to assess student mastery of 10th grade and earlier content standards. To pass the test and earn a high school diploma, students must score at least 350 on each part. The state allows students who fail either portion of the test to retake that portion in order to qualify for a diploma, we focus on student's first attempt. The original CAHSEE legislation stipulated that students in the high school class of 2003 and students in all subsequent graduating classes would be required to pass the exam in order to earn their diplomas. However, the CAHSEE was actually implemented in two steps, so that while the test was first taken by 10th graders in 2002 and 2003, it was only a graduation requirement for the cohort of 10th graders who took it in 2004. Thus the 2004 cohort is treated, while the earlier ones are controls. We have begun analysis of the exit exam data from the first of our several school districts for our analysis of the effects of the California requirement that students pass the CAHSEE.

Figure C.3 QTE estimates of CAHSEE effects on math scores



Our analyses will use official administrative data from one southern-California district to estimate the distributional effects of high stakes exit exam implementation. Preliminary analyses, reported below, focus on the effects of policy implementation on CAHSEE math test scores. Future analyses will use district enrollment and graduation data to estimate CAHSEE effects on student odds of graduating from one of the district's high schools. (Unfortunately, our administrative data do not allow us to distinguish between high school dropout and students who moved to another district.) In addition we will analyze policy effects on the distribution of student test scores on relatively low-stakes California Standards Tests and district-level tests.

We have calculated quantile treatment effects (QTE) estimates of the effect of being in a cohort required to pass the CAHSEE to graduate (treatment group) compared to being in a cohort for whom the test was low stakes (passing the exam was not required to graduate, control group). Figure C.3 above presents the estimated QTE for various points in the scaled mathematics CAHSEE score distribution along with point wise 90% confidence intervals. We find a suggestion that the exam has lead to a statistically significant increase in scores at the bottom of the distribution of math tests, but little or no effect at the top of the score distribution. We have also estimated QTE for several subgroups of students by race and ethnicity (Hispanics, Vietnamese, Whites, and others). While all of these

QTE within subgroups show something similar (mostly positive effects, larger in the middle of the distribution), there are some interesting differences as well. For example, all of the QTE for Hispanics are statistically significantly different from zero, while for both Whites and Vietnamese, those at the very top of the distribution are negative and significant. Other measures of interest and sample characteristics for this data are presented in panel 4 of Table C-1. Future analyses will also consider whether the distributional effects of CAHSEE vary between the district's 7 high schools, which might occur if schools with many students at risk of failing the exam adopted more aggressive test preparation programs.

C.3.i Progress on Specific Aim 5

Aim 5 is focused on outreach. One important component of our outreach will be presenting our work within the Network, and importantly outside the Network. We plan to present this work extensively at venues such as the annual meetings of SREE, AERA, ASA, AEA, PAA, and APPAM. We also plan to write up at least one and likely several methodological papers to illustrate these methods. One will lay out the application of QTE and the tests about subgroups for educationally-oriented audiences. The other will illustrate the application of these methods in a non-randomized setting. We will also make available Stata code and post web tutorials for each of the estimators we plan to use.

We have already completed some amount of work on dissemination of these methods. Bitler presented existing work with Hoynes on quantile regression and subgroup analysis at an interdisciplinary conference for government researchers on subgroup effects in Fall 2009. The project team presented our preliminary findings using the quasi-experimental exit exam variation of specific aim 4 to look at the effects of exit exams on the distribution of exit exam and other scores at the 2010 Society for Research on Educational Effectiveness.

C.3.j Interdependence with the Core and Other Subprojects

This project brings new methodological tools to bear on the questions of child/policy fit that motivate the Irvine Network on Interventions in Development. Traditionally, researchers have investigated issues related to child/policy fit by estimating mean effect sizes for theoretically interesting population subgroups or estimating linear interaction terms. Projects I, II, and IV propose to undertake many analyses of this sort. We approach these questions somewhat differently, estimating the effects of interventions on the distribution of continuous outcomes, or the distribution of underlying latent probabilities for dichotomous measures. We believe that these distributional methods are potentially useful for each of the Network's subprojects. In particular, we will have two contributions. First, we will disseminate our QTE and other distributional methods within the Network. Second, we will use our methods for assessing the ability of subgroup means to capture overall heterogeneity to test whether the subgroup means analyses substantially capture the important variations in outcomes across the distribution. This will particularly useful for Project I where we share a common dataset. We plan to collaborate with Project I on a paper comparing the results of distributional analyses with the results of more traditional group-based and interaction analyses, as discussed above in *C.3.e*. While the methods may not be directly applicable to the current work planned in Projects I, II, and IV, it will inform future work, and perhaps be useful in designing future interventions.

Furthermore, the Network provides us with an interdisciplinary setting in which to develop new hypotheses for distributional analysis and refine our analytical techniques. We will utilize the Network's Linux computing facilities and secure data environment in each stage of the project. In order to make the distributional tools more broadly available, we will post software and tutorials on the Network's website. These computing resources will be particularly important as we begin to develop and disseminate distributional analysis software for Stata users. More broadly, we expect to benefit from the expertise of the other members of the team in conducting and refining our analysis and in drawing conclusions about our findings. Our project is particularly interdependent with Project I: By writing a paper comparing the results of our distributional analyses with the results of the Project I's analyses of differential mean effects, we will gain new insights into the advantages and disadvantages of our approach relative to more traditional subgroup-based differential effects analyses. In addition, we will benefit from the subject-matter expertise of Duncan, Farkas, and Vandell in understanding our findings in this project. Burchinal has agreed spend 2 days a project year to help guide us in performing Aim 1. Similarly, we will draw upon the Network's advisory committee for both methodological and subject matter expertise. We anticipate that Smith's expertise in the estimation of heterogeneous program impacts will help us realize each of our aims. We will turn to Reardon and Bolt for advice regarding psychometrics and the properties of test distributions. Reardon and Loeb will help us deepen our understanding of the interventions we are evaluating and the contexts in which they operate.

D. REFERENCES CITED

Abadie, A. (2005), 'Semiparametric difference-in-difference estimators,' Review of Economic Studies 72:1, 1–19.

Abadie, A., Angrist, A., & Imbens, G. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings,' Econometrica 70:1, 91–117.

Achieve, Inc. (2007). Aligning high school graduation requirements with the real world: A road map for states. Policy Brief. Retrieved August 22, 2009 from http://www.achieve.org/files/Achieve_PolicyBrief_Dec18v4.pdf.

Allensworth, E., Nomi, T., Montgomery, N, & Lee, V. (2008). College preparatory curriculum for all: consequences of ninth grade course taking in Algebra and English on academic outcomes in Chicago. Consortium on Chicago School Research Working Paper, University of Chicago. Retrieved August 22, 2009 from http://ccsr.uchicago.edu/content/page.php?cat=3&content_id=34.

Almond, D., Hoynes, H.W., & Schanzenbach, D. (forthcoming), 'Inside the war on poverty: The impact of the Food Stamp program on birth outcomes,' Review of Economics and Statistics.

Alon, S., Domina, T., & Tienda, M. (2010). 'Stymied Mobility or Temporary Lull? The Puzzle of Lagging Hispanic College Degree Attainment.' Social Forces 88(4): 1807-1832.

Angrist, J., Chernozhukov, V., & Fernandez-Val, I. (2006), 'Quantile regression under misspecification, with an application to the U.S. wage structure,' Econometrics 73:2, 539–563.

Angrist, J., Lang, D., & Oreopoulos, P. (2009), 'Incentives and services for college achievement: Evidence from a randomized trial,' American Economic Journal: Applied Economics 1:1, 136–163.

Athey, S., & Imbens, G. (2006), 'Identification and inference in nonlinear difference-in-differences models,' Econometrica 74 2, 431–497.

Attewell, P. & Domina, T. (2008) 'Raising the bar: Curricular intensity and academic performance,' Educational Evaluation and Policy Analysis 30:1, 51–71.

Attewell, P., Lavin, D., Domina, T, & Levey, T. (2006), 'New Evidence on College Remediation,' Journal of Higher Education 77: 5.

Barnard, J., Frangakis, C.E., Hill, J.L., & Rubin, D.B., (2003), 'Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City,' Journal of the American Statistical Association 98:462, 299–311.

Barnett, W.S. (1996). Lives in the balance: age 27 cost benefit analysis of the High/Scope Perry Preschool. Ypsilanti, MI: High/Scope Press.

Bishop, J.H. (1997). 'Do curriculum-based external exit exams enhance student achievement?' Consortium for Policy Research in Education Research Report, University of Pennsylvania. Retrieved August 22, 2009 from http://www.cpre.org/images/stories/cpre_pdfs/rr40.pdf.

Bishop, J.H., Moriarty, J.Y. & Mane, F.P. (2001). 'Diplomas for learning, not for seat time: the impacts of the New York Regents examinations,' Economics of Education Review 19:4, 333–349.

Bitler, M.P., Gelbach, J.B. & Hoynes, H.W. (2006a), 'Distributional impacts of the Self-Sufficiency Project,' SRDC working paper.

Bitler, M.P., Gelbach, J.B. & Hoynes, H.W. (2006b), 'What mean impacts miss: Distributional effects of welfare reform experiments,' American Economic Review 96:4, 988–1012.

Bitler, M.P., Gelbach, J.B. & Hoynes, H.W. (2008), 'Distributional impacts of the Self-Sufficiency Project,' Journal of Public Economics 92(3–4), 748–765.

Bitler, M.P., Gelbach, J.B., & Hoynes, H.W. (2010), 'Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? Evidence from a social experiment,' unpublished.

Blundell, R., Costa Dias M., Meghir, C., & Van Reenen, J. (2004), 'Evaluating the employment impact of a mandatory job search program,' Journal of the European Economic Association 2:4, 569–606.

Buckley, J. & Schneider, M. (2002). 'What do parents want from schools? Evidence from the Internet,' Educational Evaluation and Policy Analysis 24:2, 133–144.

Campbell, F. A., & Ramey, C. T. (1995). 'Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention.' American Educational Research Journal 32:4, 743–772.

Carnoy, M. & Loeb, S. (2002), 'Does external accountability affect student outcomes? A cross-state analysis,' Educational Evaluation and Policy Analysis 24:4, 305–331.

Chaney, B., Burgdorf, K., & Atash, N. (1997). 'Influencing achievement through high school graduation requirements,' Educational Evaluation and Policy Analysis, 19:3, 229–244.

Chernozhukov, V. & Fernandez-Val, I. (2005), 'Subsampling inference on quantile regression processes,' Sankhya: The Indian Journal of Statistics 67:2, 253–76.

Chernozhukov, V., Fernandez-Val, I., & Melly, B. (2009), 'Inference on counterfactual distributions,' MIT Department of Economics Working Paper 08–16.

Chernozhukov, V., & Hansen C. (2005), 'An IV model of quantile treatment effects,' Econometrica 73:1, 245–261.

Chernozhukov, V., & Hansen, C. (2006), 'Instrumental quantile regression inference for structural and treatment effect models'. Journal of Econometrics 132:2, 491–525.

Chubb, J.E. & Moe, T.M. (1990). Politics, Markets, and America's Schools. Brookings Institute Press: Washington, D.C.

Coleman, J.S., Hoffer, T. & Kilgore, S. (1982). High school achievement: public, private, and Catholic schools compared. New York: Basic Books.

Conley, A. (2007), Patterns and Pathways: A Person-Oriented Approach to Understanding Students' Motivation to Learn, Unpublished Ph.D. Dissertation, University of Michigan.

Cunha, F., Heckman, J.J., Lochner, L., & Masterov, D. (2006), 'Interpreting the evidence on life cycle skill formation', in, E. Hanushek, and F. Welch, eds., Handbook of the Economics of Education, Volume 1 Elsevier, 697–812.

Currie, J. (2001), 'Early childhood education programs,' Journal of Economic Perspectives 15:2, 213–238.

Currie, J. & Thomas, D. (1995), 'Does Head Start make a difference?' American Economic Review 85:3, 341–364.

Currie, J. & Thomas, D. (2000), 'School quality and the longer-term effects of Head Start,' Journal of Human Resources 35:4, 755–774.

Dalton, B., Ingels, S.J., Downing, J., Bozick, R., & Owings, J. (2007). *Advanced Mathematics and Science Coursetaking in the Spring High School Senior Classes of 1982, 1992, and 2004* (NCSE 2007-312). Washington, DC : National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Dee, T.S. & Jacob, B. (2007), 'Do high school exit exams influence educational attainment or labor market performance?' in, A. Gamoran, ed., On Standards-Based Reform and the Poverty Gap: Lessons for No Child Left Behind, Washington, DC: Brookings Institution Press.

Dee, T.S. & Jacob, B. (2009), 'The impact of No Child Left Behind on student achievement.' NBER Working Paper No. 15531.

Deming, D. (2009), 'Early childhood intervention and life-cycle skill development: Evidence from Head Start,' American Economic Journal: Applied Economics 1:3, 111–134.

Domina, T. (2005), 'Leveling the Home Advantage: Assessing the Effectiveness of Parental Involvement,' Sociology of Education, 78:3, 233–249.

Domina, T. (2007), 'Higher Education Policy as Secondary School Reform: Texas High Schools After Hopwood.' Educational Evaluation and Policy Analysis 29:3, 200–217.

Domina, T. (2009), 'What Works in College Outreach: Assessing Targeted and Schoolwide Interventions for Disadvantaged Students,' Educational Evaluation and Policy Analysis 31:2, 127–152.

Eccles, J.S., Midgley, C., Wiglfield, A., Buchanan, C.M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993), 'Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families.' American Psychologist 48:2, 90–101.

Eissa, N. & Hoynes, H. (2004), 'Taxes and the labor market participation of married couples: The Earned Income Tax Credit,' Journal of Public Economics 88:9–10, 1931–1958.

Eissa, N. & Hoynes, H. (2006a), 'The hours of work response of married couples: Taxes and the Earned Income Tax Credit,' in, J. Agell and P. Sorensen, eds., Tax Policy and Labor Market Performance, MIT Press, Cambridge, MA.

Eissa, N. & Hoynes, H. W. (2006b), 'Behaviorial responses to taxes: Lessons from the EITC and labor supply,' in J. Poterba, ed., Tax Policy and the Economy, Vol. 20, MIT Press, Cambridge, MA, pp. 74– 110.

Eissa, N. & Hoynes, H. (forthcoming), 'Redistribution and tax expenditures: The Earned Income Tax Credit,' National Tax Journal.

Elacqua, G., Schneider, M. & Buckley, J. (2006.) 'School Choice in Chile: Is It Class or the Classroom?" Journal of Policy Analysis and Management 25:3, 577–601.

Evans, W. & Schwab, R. (1995) 'Finishing high school and starting college: Do Catholic schools make a difference?' Quarterly Journal of Economics 110: 941–974.

Fan, Y. & Park, S. S. (forthcoming), 'Sharp bounds on the distribution of treatment effects and their statistical inference,' Econometric Theory.

Fennema, E. (1989). 'The study of affect and mathematics: A proposed generic model for research,' in, D. B. McLeod & V. M. Adams, eds., Affect and mathematical problem solving: A new perspective (pp. 205–219). New York: Springer-Verlag.

Figlio, D. & Rouse, C. (2006), 'Do accountability and voucher threats improve low-performing schools?' Journal of Public Economics 90: 1–2, 239–255.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects,' Econometrica 75:1, 259–276.

Frangakis, C.E., & Rubin, D.B. (2002), 'Principal Stratification in Causal Inference,' Biometrics 58, 29–20.

Friedman, M. & Friedman, R. (1980). Free to Choose. New York: Harcourt Brace Jovanovich.

Gamoran, A., & Hannigan, E.C. (2000) 'Algebra for Everyone? Benefits of College Preparatory Mathematics for Students with Diverse Abilities in Early Secondary School.' Educational Evaluation and Policy Analysis, 22, 241–254.

Garces, E., Thomas, D., & Currie, J. (2002), 'Longer-term effects of Head Start,' American Economic Review 92:4, 999–1012.

Giedd, J. (2008). 'The teen brain: insights from neuroimaging.' Journal of Adolescent Health, 42(4): 335–343.

Grodsky, E., Warren, J. & Kalogrides, D. (2009), 'State High School Exit Examinations and NAEP Long-Term Trends in Reading and Mathematics, 1971–2004.' Educational Policy 23: 589–614.

Hahn, J., Todd, P., and van Der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design,' Econometrica 69:1, 201–209.

Hanushek, E. & Raymond, M. (2004), 'The effect of school accountability systems on the level and distribution of student achievement,' Journal of the European Economic Association 2: 2–3, 406–415.

Hastings, J., Kane, T., & Staiger, D. (2005), 'Parental preferences and school competition: evidence from a public school choice program.' National Bureau of Economic Research Working Paper No.11805.

Heckman, J.J. (2006), 'Skill formation and the economics of investing in disadvantaged children,' Science 312:5782, 1900–1902.

Heckman, J.J. (2007), 'The productivity argument for investing in young children.' *National Bureau of Economic Research Working Paper No. W13016,* April 2006.

Heckman, J.J., Ichimura, H., Smith, J. & Todd, P. (1998), 'Characterizing selection bias using experimental data,' Econometrica 66:5, 1017–98.

Heckman, J.J., Ichimura, H., & Todd, P. (1998), 'Matching as an econometric evaluation estimator,' Review of Economic Studies 65, 261–294.

Heckman, J. J., Smith, J. & Clements, N. (1997), 'Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts,' Review of Economic Studies 64, 487–535.

Hochshild, J. & Scrovronick, N. (2003), The American Dream and the Public Schools. New York: Oxford University Press.

Hoffer, T., Greeley, A.M. & Coleman, J.S. (1985), 'Achievement growth in public and Catholic schools,' Sociology of Education, 58: 74–97.

Holland, P. (1986), 'Statistics and causal inference,' Journal of the American Statistical Association 81:396, 945–970.

Howell, W.G., Wolf, P.J., Campbell, D.E., & Peterson, P.E., (2002), 'School vouchers and academic

performance: Results from three randomized field trials,' Journal of Policy Analysis and Management 21:2, 191–217.

Hoynes, H. (1996), 'Welfare transfers in two-parent families: Labor supply and welfare participation under the AFDC-UP program,' Econometrica 64:2, 295–332.

Hoynes, H., (1997), 'Does welfare play any role in female headship decisions?' Journal of Public Economics 65:2, 89–117.

Hoynes, H. (2000), 'Local labor markets and welfare spells: Do demand conditions matter?' Review of Economics and Statistics 82:3, 351–368.

Hoynes, H., & Schanzenbach, D. (2009), 'Consumption responses to in-kind transfers: Evidence from the introduction of the food stamp program,' American Economic Journal: Applied Econonomics.

Human Resources Research Organization (2008), Independent evaluation of the California High School Exit Examination (CAHSEE) fourth biennial report.

Jacob, B., (2001), 'Getting tough? The impact of high school graduation exams,' Educational Evaluation and Policy Analysis 23: 2, 99–121.

Knudsen, E. I., Heckman, J. J., Cameron, J. L., & Shonkoff, J. P. (2006), 'Economic, neurobiological, and behavioral perspectives on building America's future workforce.' Proceedings of the National Academy of Sciences 103:27, 10155–10162.

Kordas, G. (2006), 'Smoothed binary regression quantiles,' Journal of Applied Econometrics 21:3, 387–407.

Koenker, R. & Bassett, G. (1978), 'Regression quantiles,' Econometrica 46, 33–50.

Koenker, R. & Bilias, Y. (2001), 'Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments,' Empirical Economics 26:1, 199–220.

Krueger, A. & Zhu, P. (2004a), 'Another look at the New York City school voucher experiment,' American Behavioral Scientist 47:5, 658–698.

Krueger, A. & Zhu, P. (2004b), 'Inefficiency, subsample selection bias, and nonrobustness,' American Behavioral Scientist 47:5, 718–728.

Loveless, T. (1999), *The Tracking Wars: State Reform Meets State Policy*. Washington, DC: Brookings Institution Press.

Lucas, S. (1999), *Tracking Inequality: Stratification and Mobility in American High Schools*. New York: Teachers College Press.

Ludwig, J. & Miller, D. (2007), 'Does Head Start improve children's life changes? Evidence from a regression discontinuity approach,' Quarterly Journal of Economics 122: 1, 159–208.

Magnuson, K., Meyers, M., Ruhm, C. & Waldfogel, J., (2004), 'Inequality in Preschool Education and School Readiness,' American Educational Research Journal 41:1, 115–157.

Martorell, F., (2004), 'Do High School Graduation Exams Matter? A Regression Discontinuity Approach.' Working paper, University of California, Berkeley.

Mayer, D.P., Peterson P.E., Myers, D.E., Tuttle, C.C., & Howell, W.G. (2002), 'School choice in New York City after three years: An evaluation of the school choice scholarship program.' Mathematica Policy Research, Inc.: Washington, D.C.

Morgan, S.L., (2001), 'Counterfactuals, causal effect hetereogeneity, and the Catholic school effect on learning,' Sociology of Education 74, 341–374.

National Commission on Excellence in Education, (1983), A Nation At Risk: The Imperative for Education Reform. Washington, DC: U.S. Department of Education. Retrieved August 22, 2009 from http://www.ed.gov/pubs/NatAtRisk/index.html.

National Diploma Project, (2004), Ready or Not: Creating a High School Diploma That Counts. Washington, DC: National Diploma Project. Retrieved August 22, 2009 from http://www.achieve.org/files/ADPreport_7.pdf.

National Governor's Association, (2005), An Action Agenda for Improving America's High Schools. Retrieved August 22, 2009 from http://www.nga.org/Files/pdf/0502actionagenda.pdf.

Neal, D. & Schanzenbach, D. (2010), 'Left behind by design: Proficiency counts and test-based accountability,' Review of Economics and Statistics 92: 2, 263-283.

NICHD Early Child Care Research Network, (2003), 'Modeling the impacts of child care quality on children's

preschool cognitive development,' Child Development 74, 1454–1475.

Oakes, J., (1985). Keeping Track: How Schools Structure Inequality. New Haven: Yale University Press.

Papay, J.P., Murnane, R.H. & Willet, J.B., (2008), 'The consequences of high school exit examinations for struggling low-income students: Evidence from Massachusetts.' National Bureau of Economic Research Working Paper No. W14186.

Penner, A.M., (2008), 'Gender differences in extreme mathematics achievement: An international perspective on biological and social factors.' American Journal of Sociology 114:S1, S138–70.

Penner, A.M. & Paret, M. (2008), 'Gender differences in mathematics achievement: Exploring the early grades and the extremes.' Social Science Research 37:1, 239–253.

Peterson, P.E., & Howell, W.G., (2004), 'Efficiency, bias, and classification schemes: A response to Alan B. Krueger and Pei Zhu,' American Behavioral Scientist 47:5, 699–717.

Pintrich, P., Conley, A., & Kempler, T. (2003), 'Current issues in achievement goal theory and research,' International Journal of Educational Research 39, 319–337.

Powell, A.G., Farrar, E. & Cohen, D.K., (1985), The Shopping Mall High School: Winners and Losers in the Educational Marketplace. Boston: Houghton Mifflin.

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005), 'Head Start Impact Study: First Year Findings,' prepared for U.S. Department of Health and Human Services, Administration for Children and Families: Washington DC.

Reardon, S., Atteberry, A., Arshan, N., & Kurlaender, M. (2009). Effects of the CaliforniaHigh School Exit Exam on Student Persistence and Graduation. Institute for Research on Educational Policy and Practice Working Paper, Stanford University. Retrieved August 22, 2009 from http://www.stanford.edu/group/irepp/cgi-bin/joomla/docman/reardon-et-al-cahsee-study/download.html.

Reback, R. (2008), 'Teaching to the rating: School accountability and the distribution of student achievement,' Journal of Public Economics 92, 1394–1415.

Rosenbaum, J.E., (1999), 'If tracking is bad, is detracking better? A study of a detracked high school,' American Educator.

Rosin, M., Barondess, H., & Leichty, J. (2009), 'Algebra policy in California: great expectations and serious challenges.' EdSource Report.

Rubin, D.B., (1974), 'Estimating causal effects of treatments in randomized and non-randomized studies,' Journal of Educational Psychology 66, 688–701.

Schoenfeld, A. (1992). 'Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics.' Handbook of research on mathematics teaching and learning, 334–370.

Schweinhart, L.J., Barnes, H.V., & Weikart, D.P. (1993). Significant benefits: The High/Scope Perry Preschool Study through age 27. Ypsilanti, MI: High/Scope Educational Research Foundation.

Stanovich, K. (1986), 'Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy,' Reading Research Quarterly 22:4, 360–407.

United States Department of Health and Human Services, Office of Head Start. (2008), http:www.acf.hhs.gov/programs/ohs/about/fy2008.html (accessed August 2009).

van Der Klaaw, W. (2002), 'Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach,' International Economic Review 43:4, 1249–1287.

Warren, J.R. & Edwards, M.R., (2005), 'High school exit examinations and high school completion: Evidence from the early 1990s.' Educational Evaluation and Policy Analysis 27:1, 53–74.

Warren, J. R., Jenkins, K. N., & Kulick, R. B., (2006), 'High school exit examinations and state-level completion and GED rates, 1975–2002.' Educational Evaluation and Policy Analysis 28:2, 131–152.

Weiher, G.R. & Tedin, K. (2002). 'Does Choice Lead to Racially Stratified Schools? Charter Schools and Household Preferences.' Journal of Policy Analysis and Management, 21:1, 79–92.

Wong, M., Cook, T.D. & Steiner, P.M. (2010). 'No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series." Working Paper, Northwestern University Institute for Policy Research. Retrieved September 15, 2010 from: http://www.northwestern.edu/ipr/publications/papers/2009/wp0911.pdf.

E. PROTECTION OF HUMAN SUBJECTS

We are planning to use 4 separate datasets: the Head Start Impact Study (HSIS), the School Choice Scholarships Program Evaluation (SCSPE), the California Motivation Project-Math (CAMP-Math), and administrative data from several school districts on high school students. The second (School Choice Scholarships Program Evaluation data for specific aim 2), third (CAMP-Math), and fourth of the datasets (the school administrative data for specific aim 4) will only be analyzed as secondary data. The first dataset is as yet unavailable publicly although promised soon from NCES. Previous experience with NCES suggests it will also be secondary data.

E.1 Protection of Human Subjects in the Head Start Impact Study

E.1.a Risks to Human Subjects

E.1.a.1 Human subjects involvement, characteristics, and design. The first dataset we propose to use is the Head Start Impact Study (HSIS) public use data, which is not yet available, but will be available for secondary analyses as a public use dataset. Based on information from the public first year report, the participants for this study are the 4,667 participants in the HSIS. This study is a randomized control trial of the effects of Head Start on outcomes for children aged 3 or 4 who applied to participate in Head Start. The goal was to get nationally representative measures of the effects of Head Start. First, centers were randomly selected to be in the study, selected to be nationally representative of centers with waiting lists. This led to a sample of 84 nationally representative grantees/delegate agencies with waiting lists. Then, children from the selected centers who were on the waiting lists were randomized into one of two arms in the study, the treatment group who got into the Head Start program for which they were on the waiting list, and the control group, who did not and could obtain non Head Start services in their community. Head Start is a federal grant program which makes direct grants to Head Start centers. Children are eligible if they are low income (under poverty) or public assistance recipients. As a result, the participants in the HSIS are representative of the population of 3 and 4 year old children who applied for the first time to Head Start at centers where there were waiting lists, and where the Head Start program was not migrant/tribal or Early Head Start only. The eventual sample consisted of 4,667 3 and 4 year olds in the 2002–2003 school year, with 2783 treatment group members and 1884 control group members. The study purposely oversampled the treatment group because this increased the number of centers which could be included (centers which would have space for more children could be included). Because of the goal of the study, the data are only representative of children who apply to Head Start. Thus, high income children and other children who may not apply to Head Start are not eligible for inclusion. We will use data on all children available in the public use dataset.

E.1.a.2 Sources of materials. The HSIS data were collected by Westat with the help of subcontractors. We will be performing secondary analyses of previously collected data, and thus will not be in possession of any materials other than the public use dataset provided. Demographic, socioeconomic, cognitive, achievement, and behavioral information will be included in the data, from direct child assessments, parental reports, and teacher reports. All will be linked with a unique subject identifier only connected to personally identifiable information by the contractors. None of this information will be available to us, and as far as we know, no geographic identifiers below the state level will be available to us. However, as the HSIS data are not yet publicly available, there is some possibility that these data might contain more precise geographic identifiers. As such, we plan to analyze the data as outlined below.

E.1.a.3 Potential risks. The issues of informed consent and risk to the participants were dealt with by Westat during the original data collection. All parents were informed at the time of recruitment that a study was going to be conducted that might affect who received Head Start, especially for the control group children. We are only proposing to use secondary data here. Thus, the major risk associated with these secondary analyses of the HSIS described in the current proposal Aim 1 involves data security and threats to subject confidentiality associated with deductive disclosure of respondent identity.

E.1.b Adequacy of Protection Against Risks

E.1.b.1 Recruitment and informed consent. We will have no contact with research subjects and will play no role in recruitment or consent.

E.1.b.2 Protections against risk. No items will be released in the public use HSIS data that would allow identification of specific families. Thus, geographic data below the state level and other sensitive data will be removed from the data. The public use data have been stripped of names and addresses and will only be identified according to subject ID numbers. We will have no access to the data linking these IDs to individuals.

In addition, the following data protection protocol will be followed stringently by our team of investigators and research assistants.

(1) All the data will be analyzed only on LINUX-based accounts with carefully protected accounts and passwords.

(2) Analyses will be conducted at an aggregate level rather than at a level of detail that would allow identification of individual subjects.

E.1.c Potential Benefits of the Proposed Research to Human Subjects and Others

There will be no direct immediate benefits to participants of this research. However, the knowledge gained from this study has the potential to contribute to policy about expanding Head Start and changing the goals of the program. The potential benefit, while distant, outweighs the minimal risks associated with the conduct of secondary data analysis of the HSIS.

E.1.d Importance of the Knowledge to be Gained

The potential knowledge gained by this study is great, given the fact that it will use the first large scale nationally representative random assignment study of Head Start. Not only were the participants randomly assigned, but also the sites themselves were randomly assigned. Our analysis of effects across the distribution will add to our understanding of which parts of the achievement or socio-behavioral distributions gain from Head Start. Our analysis of whether subgroups predicted to identify groups that would differentially respond to Head Start captures the heterogeneity in the effects across the distribution informs discussion about targeting. These findings will inform how Head Start impacts are valued and will help disseminate these distributional methods to interested researchers in education.

E.2 Protection of Human Subjects in the School Choice Scholarships Program Evaluation

E.2.a Risks to Human Subjects

E.2.a.1 Human subjects involvement, characteristics, and design. The second dataset we use is the New York City School Choice Scholarships Program (NYCSCSP) restricted-use data. These data were collected by Mathematica Policy Research, Inc. for the purpose of evaluating the effects of the Scholarships program. The data were collected as part of a randomized control trial evaluating the effect of offering vouchers to attend private school worth $1400 to low income children in grades K–4 in New York City schools in 1997. The 2080 participants with outcome data were selected from a list of 10,000 eligible applicants (children had to be income eligible for the School Lunch program and be attending a public school). Students at schools with scores below the city median were given a higher number of slots, but weights were created to make data generalizable to all eligible applicants. As a result, the participants in the NYCSCSP are representative of the population of school lunch eligible children attending public schools in New York City in 1997 who wanted a voucher. Because of the goal of the study, the data are only representative of children who applied for vouchers, were in grades K–4, and were school lunch eligible in New York City public schools. Thus, high income children and other children who are not eligible for School Lunch or not attending public schools or in grades 5 or above were not eligible for inclusion. We will use data for all of the children for whom data are available in the restricted use dataset.

E.2.a.2 Sources of materials. The NYCSCSP data were collected by Mathematica Policy Research, Inc. We are performing secondary analyses of previously collected data, and thus will not be in possession of any materials other than the restricted use dataset provided. Demographic, socioeconomic, cognitive, achievement, and behavioral data will be included in the data, from direct child assessments, parental reports, and teacher reports. All will be linked with a unique subject identifier only connected to personally identifiable information by Mathematica. None of this information will be available to us, and as far as we know, no geographic identifiers will be available to us beyond city of residence in 1997.

E.2.a.3 Potential risks. The issues of informed consent and risk to the participants were dealt with by Mathematica during the original data collection. We are only proposing to use secondary data here. Thus, the major risk associated with these secondary analyses of the NYCSCSP described in the current proposal Aim 2 involves data security and threats to subject confidentiality associated with deductive disclosure of respondent identity.

E.2.b Adequacy of Protection Against Risks

E.2.b.1 Recruitment and informed consent. We will have no contact with research subjects and will play no role in recruitment or consent.

E.2.b.2 Protections against risk. We have obtained access to the restricted use data. No items are in the restricted use NYCSCSP data that would allow identification of specific families. Thus, addresses and other sensitive data are not present in the data we received. The restricted use data have been stripped of names and addresses and will only be identified according to subject ID numbers. We will have no access to the data linking these IDs to individuals. In addition, the following data protection protocol will be followed stringently by our team of investigators and research assistants.

(1) The application to Mathematica to obtain the data included a data security plan and a list of authorized data users. All authorized users will be required to execute a confidentiality agreement supplied by Mathematica when they approve our application to use the data.

(2) All the data will be analyzed only on a stand alone LINUX machine not connected to the Internet via LINUX-based accounts with carefully protected accounts and passwords. The analysis data will be kept in a directory that is only accessible to members of the team of authorized data users. The original CD will be kept in a locked cabinet.

(3) Analyses will be conducted at an aggregate level rather than at a level of detail that would allow identification of individual subjects.

E.2.c Potential Benefits of the Proposed Research to Human Subjects and Others

There will be no direct immediate benefits to participants of this research. However, the knowledge gained from this study has the potential to contribute to policy about expanding school choice via vouchers or other methods. The potential benefit, while distant, outweighs the minimal risks associated with the conduct of secondary data analysis of the SCPSE.

E.2.d Importance of the Knowledge to be Gained

The potential knowledge gained by this study is substantial. Our analysis of effects of school vouchers across the test score distribution will add to our understanding of which parts of the achievement distribution for children seeking to leave public schools gain from vouchers. Our analysis of whether subgroups predicted to identify groups that would differentially respond to an offer of a voucher captures the heterogeneity in the effects across the distribution informs discussions about targeting. These findings will inform policy debate about school choice and will help disseminate these distributional methods to interested researchers in education.

E.3 Protection of Human Subjects in the California Motivation Project-Math

E.3.a Risks to Human Subjects

E.3.a.1 Human subjects involvement, characteristics, and design. The third dataset that we propose to use is the California Motivation Project-Math (CAMP-Math). We are proposing secondary analysis of existing data for the entire CAMP-Math sample (N=13,829). The CAMP-Math database provides detailed assessments of student mathematics motivation in the fall and spring of two consecutive school years (2004–5 and 2005–6) for 13,829 students in 14 low-income schools (7 middle, 7 high) in 4 large urban school districts. Administrative data provide information on children's test scores both before and after the years in which the motivation surveys were administered. It covers grades 7 through 12 with about 500 classrooms at each wave. The schools served ethnically (69% Latino; 16% Asian, primarily Vietnamese; and 11% Caucasians) and linguistically (40% English learners) diverse learners and a majority of the students were eligible for free or reduced lunch. The county from which the sample was drawn has a foreign-born rate of 29.9% (US Census Bureau, 2009, based on 2000 census), and the percent of immigrants was even higher in the sample schools. Only students present in a math class in the participating schools when the surveys were administered are included in the database. No other exclusion criteria exist for inclusion in the CAMP-Math sample. Because this was a study of school-age motivation and achievement, it was necessary to include children in the study.

E.3.a.2 Sources of materials. The existing CAMP-Math dataset includes administrative data provided by the school districts, achievement test scores, and four waves of motivation surveys. The version of the data we will have access to contains no personal identifiers.

E.3.a.3 Potential risks. The only risk associated with the analyses of existing data described in this proposal involves the potential for a breach of confidentiality.

E.3.b Adequacy of Protection Against Risks

E.3.b.1 Recruitment and informed consent. Parental consent for participation in the original CAMP-Math project was obtained by participating schools with letters to parents in English, Spanish, and Vietnamese.

E.3.b.2 Protections against risk. For secondary analysis of existing data, we will protect against risk by adhering to a strict data security plan. Access to the data by our investigators and research assistants will require completion of human subjects training and certification by UC Irvine's Institutional Review Board. All analytic data files have been stripped of participant names and addresses, and are identified only according to subject ID numbers. Participant names and addresses are stored apart from motivation and achievement records and are not available to us.

The following data protection protocol will be followed stringently by our team of investigators and research assistants in using the analytic files (which have been stripped of the confidential information).

(1) As noted above, we will require investigators and research assistants to complete human subjects training and certification.

(2) All the data will be analyzed only on LINUX-based accounts with carefully protected accounts and passwords. The data will be kept in a directory that is only accessible to members of the team of authorized data users.

(3) Analyses will be conducted at an aggregate level rather than at a level of detail that would allow identification of individual subjects.

E.3.c Potential Benefits of the Proposed Research to Human Subjects and Others

In the long term, the knowledge gained from this study has the potential to contribute to educational policy reforms that may directly benefit the achievement and quality of life of participants. This potential benefit, while very distant, outweighs the minimal risk associated with the conduct of the secondary data analyses proposed in this study as well as the proposed collection of survey data.

E.3.d Importance of the Knowledge to be Gained

This study will contribute to our understanding of the distributional effects of curricular policy on student motivation and achievement. These societal benefits are far greater than the minimal risks to subjects noted above. and will help disseminate these distributional methods to interested researchers in education. While, as noted above, these positive outcomes are quite far in the future, the minimal risk in this project is far outweighed by the potential gains to be realized.

E.4 Inclusion of Women and Minorities

The HSIS study will include only low income children who apply to Head Start as only these children are eligible for Head Start. To the extent that women have children who apply to Head Start, they will be represented in the data about parenting and parental assessment of children's sociobehavioral outcomes. Because all groups of Head Start children except migrant/American Indian only program applicants or applicants to programs where there was no waiting list, the set of children is nationally representative of the set of children applying to Head Start. Thus, groups from all race/ethnic minorities will be represented to the extent they are in the relevant population of centers, including Whites, Blacks, Hispanics, Asians/Pacific Islanders, and others. The goal of the experiment is to assess the effects of Head Start, thus only children of the age to be eligible for Head Start and who applied to it are included. Children are followed for several years after their applications to Head Start, so our longitudinal data will include children from ages 3 through 6 at least.

The NYCSCSP study will include only low income children in grades K–4 in public schools in New York City in 1997 who applied for vouchers. To the extent that women are the mothers of this set of children, they will be represented in the data about parental assessment of schools and satisfaction with the children's schools. Because the sample is a subset of low income children in public schools in New York City whose parents wanted school vouchers, the set of children in the data are representative of this larger population of children. Thus, groups from all race/ethnic minorities were represented but in proportion to their representation among the eligible universe. Race/ethnicity is reported based on the race/ethnicity of the mother/female guardian in the final report (Puma et al., 2005). Thus, the bulk of the mothers/female guardians were black, Puerto Ricans, and other non-Puerto Rican Hispanics. Since the goal of the experiment was to study the effects of school choice for these children in grades K–4, only children of the age range enrolled in those grades are included. There are 3 years of follow up data on the children, so our longitudinal data will include children from ages about 5–9.

The CAMP-Math data we will use for Aim 3 will include all children enrolled in 8th grade between 2002 and 2008 in CAMP-Math schools. These include schools from three districts. Since these are data on children only, women are excluded. To the extent that these children differ from all children, this will be reflected in our sample. These schools are disproportionately low income, and the students are disproportionately Hispanic

and Vietnamese. But this reflects the public school enrollment in this area among schools who participated in the Camp-Math data. Children from all minority groups are included according to their representation in the CAMP-Math Schools.

The administrative data we will use for Aim 4 will include all children in the schools in our sample (which come from three districts) who were eligible to take the CAHSEE exit exam in academic years 2003–04, 2004–05, and 2005–06. Since they are data only on children, women are excluded. Children of all minority groups are included to the extent they are represented in these districts.

E.5 Targeted/Planned Enrollment Tables

Targeted/planned enrollment tables for the data we will use are appended to this research plan.

E.6 Inclusion of Children

The HSIS study is a study of the effects of Head Start enrollment on children aged 3 or 4 at first Head Start application. Thus, data are included for a nationally representative children applying to Head Start in areas where there is excess demand. Head Start eligibility is conditioned on being low income or a public assistance recipient, so the data includes only such children. Children of all races and ethnicities are included to the extent that they are in the sample universe. Since Head Start is only for children aged 3 and 4, data are included for children entering Head Start at ages 3 or 4, and following these children through 1st grade.

The NYCSCSP study is a study of the effects of an offer of school vouchers on School Lunch eligible children in New York City schools in grades K–4 applying to get vouchers on achievement and other outcomes. Thus, data are included for a representative sample of the set of children in this universe. To the extent that this universe differs from the full sample of children in grades K–4 nationally, the samples will differ. Children were followed for 3 years, so children from the age groups to be in grades K–7 will be represented in our data.

The CAMP-Math data we will use for Aim 3 will include all children enrolled in 8th grade between 2002 and 2008 in CAMP-Math schools. These include schools from three districts. To the extent that these children differ from all children, this will be reflected in our sample. These schools are disproportionately low income, and the students are disproportionately Hispanic and Vietnamese. But this reflects the public school enrollment in this area among schools who participated in the Camp-Math data. Children from all minority groups are included according to their representation in the CAMP-Math Schools. Since our goal is to examine the effects of being encouraged to take Algebra 1 in 8th grade on 8th grade motivation and achievement and later academic outcomes, we only include data for 8th graders (and data for these same children later on).

The administrative data we will use for Aim 4 will include all children in the schools in our sample (which come from three districts) who were eligible to take the CAHSEE exit exam in academic years 2003–04, 2004–05, and 2005–06. Children from all minority groups are included according to their representation in the schools from three districts in our sample. Since our goal is to look at the effects of exit exams on achievement and later outcomes, we only include data for 10th graders (and data for these same children later on).

F. VERTEBRATE ANIMALS

Not applicable

G. SELECT AGENT RESEARCH

Not applicable

H. INVESTIGATORS

Marianne Bitler, Associate Professor in the Department of Economics at the University of California, Irvine (UCI), is the PI of this subproject. Of particular relevance for this subproject is Bitler's expertise in program evaluation and the use of explicitly distributional methods for looking at the effects of interventions. Bitler's research focuses on the effects of social insurance programs on disadvantaged populations and on the economics of the family. Together with Hilary Hoynes and Jonah Gelbach, Bitler has had a long and productive collaboration on the topic of recent welfare reforms. Recently, Bitler, Gelbach and Hoynes have pushed their work on welfare beyond the mean treatment effects to look at effects across the full distribution of outcomes. Based on this expertise in program evaluation, Bitler was invited to present a lecture about Subgroups and Quantile Treatment Effects at a September 2009 meeting on Subgroup Analysis in Prevention and Intervention Research, sponsored by the Office of Planning, Research and Evaluation in the Department of Health and Human Services' Administration for Children and Families in partnership with the Office of the Assistant Secretary for Planning and Evaluation, the Centers for Disease Control, the National Institute on Drug Abuse, the National Institute on Mental Health and the Department of Education's Institute of Education Sciences.

Two papers have come out of Bitler's work using distributional methods to examine the effects of welfare reforms (Bitler, Gelbach, & Hoynes 2006b, 2008). The papers show that looking at distributions, rather than calculating means or means within particular subgroups, provides a better understanding of treatment heterogeneity. Using experimental data from two welfare reform experiments, they examine the effects of these reforms across the distribution of income, earnings, and transfers. In both settings, there were predictions that individuals should be differentially affected by the reforms depending on their likely outcomes in the absence of reform. For example, the first experiment implemented in Connecticut allowed women in the treatment group to keep all earnings while on welfare until they reach the poverty limit, while the control group faced the usual AFDC policy of a 100% tax rate on earnings. Economic theory gives clear predictions that some women with relatively high hours/earnings under AFDC assignment should be likely to reduce hours and earnings under assignment to the treatment group, while some women with zero hours of work under AFDC assignment should increase their hours of work if assigned to the treatment group. This leads to a prediction that some parts of the counterfactual earnings distribution under AFDC would gain and some would lose under treatment group assignment, leading to a possibly small mean effect which obscured large effects at different points in the distribution. In both studies, quantile treatment effects showed effects that were zero at the bottom of the earnings distribution, then positive and then zero or negative at the top, consistent with the theoretical predictions. In a recent working paper, Bitler, Gelbach and Hoynes developed ways to test whether this kind of heterogeneity can be explained simply by different mean effects within subgroup (Bitler, Gelbach, & Hoynes, 2010). Together with other project investigators, Bitler and Hoynes plan to bring these methods to bear on educational data, where much of the existing work fails to consider effects on distributions and considers treatment effect heterogeneity only in the context of estimating means within subgroups such as the set of students with high or low test scores before an intervention.

AnneMarie Conley is Assistant Professor in the Department of Education, UCI. Her subject-matter expertise in motivation, especially with regard to achievement goal theory is particularly relevant to this subproject (Conley 2007; Pintrich, Conley & Kempler 2003). Conley is trained as an educational and developmental psychologist, with a focus on motivation and its influence on achievement during the middle and high school years. Conley designed and directed the California Motivation Project (CAMP-Math). The multiple waves of motivation data from 14,000 CAMP students will support Aim 3 of this project, and her relationship with the school districts has enabled access to the data to be used in Aim 4 and used below for preliminary studies. She is also co-PI on a new $2 million grant from the National Science Foundation (NSF) with the school districts whose data we will use in Aims 3 and 4. Her responsibility for data analysis on the NSF grant is a synergistic activity that will maximize the student-level data available now and in the future for distributional analysis of exit exam (Aim 4) and Algebra course-taking (Aim 3) data.

Thurston Domina is Assistant Professor in the Department of Education, UCI. Domina brings broad educational policy expertise to this project, as well as methodological strengths related to the quasi-experimental estimation of program effects. His research interests span the preschool-to-higher education continuum, focusing on the ways in which policy effects differ across racial and socio-economic subgroups. He has studied elementary school parental involvement initiatives (Domina 2005); educational outreach and information programs for high school students (Domina 2007; Domina 2009); standards and accountability policies designed to raise the academic intensity of high school student coursework (Attewell & Domina 2008; Domina 2007); affirmative action and other programs related to college admissions (Alon, Domina, & Tienda 2010; Domina 2007); and remedial education programs for college students (Attewell, Lavin, Domina, & Levey 2006).

Hilary Hoynes is a Professor in the Department of Economics at the University of California, Davis; co-editor of the American Economic Journal: Economic Policy; research associate of the National Bureau of Economic Research; and senior research affiliate of the National Poverty Center at the University of Michigan. Hoynes specializes in low-wage labor markets and the study of tax and transfer programs for poor families. Her work looks at the effects of tax and transfer programs on labor supply, family formation, and poverty. Her published work on U.S. cash welfare programs includes Hoynes (1996), on the effect of welfare for two-parent families on labor supply and welfare participation; Hoynes (1997) on the effect of welfare on the probability of being a female head of household; and Hoynes (2000), on the effect of local labor markets on welfare participation. She is an expert on the effects of the Earned Income Tax Credit and has published several papers about the effects of the EITC on the extensive and intensive margins of labor supply (Eissa & Hoynes 2004, 2006a, 2006b, forthcoming). She has a comprehensive research program on the economic and health benefits

of the food stamp program (e.g., Hoynes & Schanzenbach 2009; Almond, Hoynes, & Schanzenbach forthcoming). Hoynes' experience with NIH includes being on a review panel for demography center grants and giving a talk on "Income Support Policies and Disparities in Health" at the Fall 2006 NIH Conference on Understanding and Reducing Disparities in Health as well as receipt of R01 funding.

Andrew Penner is Assistant Professor of Sociology, UCI. Of importance for this project, Penner's research seeks to bring a distributional perspective to education research, and focuses on gender differences in mathematics achievement. He has used quantile regression to examine how gender differences in mathematics achievement among secondary students vary across the distribution in 22 countries, documenting that patterns of gender differences in mathematics achievement vary widely across countries (Penner 2008). He has also used these methods to investigate the emergence of gender differences in the mathematics achievement of elementary school students in the United States, showing that gender differences at the top of the distribution exist even in the fall of kindergarten (Penner & Paret 2008).

Margaret Burchinal is a Professor in the Department of Education at the University of California, Irvine. Prior to joining the Irvine faculty in 2007, Burchinal was a Senior Scientist and Director of the Data Management and Analysis Unit at the Frank Porter Graham Child Development Institute and Research Professor of Psychology at the University of North Carolina at Chapel Hill. Burchinal is an applied statistician with broad experience on developmental projects, especially in child care research. Burchinal is a prolific researcher, with a large body of work showing the importance of quality child care in reducing race and economic gaps at school entry. As an applied methodologist, she has helped to bring many techniques into use in by developmental and educational researchers. Her expertise in these topics and experience as a member of the Advisory Board for the National Head Start Impact Study will enable her to provide expert advice about developmental theory and measurement to be used in Aim 1.

I. CONSORTIUM/CONTRACTUAL ARRANGEMENTS

The bulk of the key personnel on our subproject (Bitler, Burchinal, Conley, Domina, and Penner) are faculty at UCI. Thus, we will have an easy time meeting and coordinating our research, and will have frequent additional interactions as part of the Network interactions. Co-Investigator Hilary Hoynes has worked productively in the past with Bitler. She will perform any data analysis via accounts on the LINUX servers supplied by the Network. She has budgeted one trip per year to UCI to coordinate issues that cannot be worked out on the phone. Bitler and Hoynes interact frequently at professional conferences as well.

J. LETTERS OF SUPPORT

A letter of collaboration from Co-Investigator Hilary Hoynes is appended to this research plan.

K. RESOURCE SHARING PLAN

Our program project will analyze a number of datasets and will use some data (CAMP-Math) previously collected by AnneMarie Conley, a co-Investigator on this project. The data used for Aim 1, the HSIS data, will be public use data available from NCES shortly. The data used for Aim 2, the NYCSCPC data, is available upon completion of an application process to Mathematica Policy Research, Inc. The data used for Aim 3 (CAMP-Math) are discussed below. Aim 4 of our project will involve analysis of secondary data (administrative data) provided by three school districts. Aggregate school level data are available at the State of California's Department of Education website. The districts in question have policies for use of their individual-level data.

Aim 3 uses an existing database (CAMP-Math). To enable outside researchers to access these data, we intend to negotiate an agreement with the University of Michigan's ICPSR that would release these data under special contractual conditions. These conditions are consistent with ICPSR's suggested protocols and include: i) investigators must be affiliated with an institution of higher education, a research organization, or a government agency; ii) investigator must submit curriculum vitae; detailed information on how restricted data will be stored, i.e. location of computers where data will be stored, who will have access to them, and how it will be ascertained that access is limited; an agreement to abide by the limitation that the data will be used solely for research, and that no attempt will be made to identify specific individuals, families, households, schools, or institutions will be made (the CAMP-Math data that are released do not contain names of respondents); iii) a requirement that, if the identity of any person, family, or household should be discovered inadvertently, then (a) no use will be made of this knowledge; (b) ICPSR will be advised of the incident on a timely basis; (c) the information that would identify the person, family, or household will be safeguarded or destroyed as requested by ICPSR; and (d) no one else will be informed of the discovered identity; iv) that data published or in any other way released should never permit a determination of the identity of any individual either alone or when

used in combination with other known data; v) that no attempt will be made to link this restricted data with any other dataset, including other datasets provided by ICPSR, unless specifically identified in the approved IRB proposal and specifically approved in writing by ICPSR. Moreover, we will require that investigators submit annual IRB renewal approvals to ICPSR.