

The Effects of Welfare Reform on the Academic Performance of Children in Low-Income Households

Amalia R. Miller^{*} and Lei Zhang^{†‡}

June 2008

Abstract

During the 1990s, U.S. welfare policy underwent dramatic reforms aimed at promoting employment and reducing dependence. Although the immediate effects on adult labor supply and family income have been studied extensively, this paper is the first to evaluate the long-run effects on children's well-being. Using a decade of national math achievement data, and controlling for contemporaneous changes in education policy and environment, we associate welfare reform with relative improvements in low-income students' test scores. Larger gains occur in states with greater initial welfare caseloads and caseload reductions. Preliminary analysis shows relative improvements in low-income children's time use and parental interaction.

^{*} Department of Economics, University of Virginia, Charlottesville, VA, armiller@virginia.edu.

[†] Department of Economics, Clemson University, Clemson, SC, zlei@clemson.edu.

[‡] We appreciate comments from Doug Bernheim, William Dougan, Eric Hanushek, Caroline Hoxby, John Pepper, Macke Raymond, Sarah Turner, seminar participants at Clemson, Maryland, Virginia, Stanford, UC-Merced, San Francisco State, and the BLS, and conference participants at the AEFA annual meetings, NBER Education Program spring 2007 meetings, 2007 SOLE meetings, and 2007 Econometric Society North American summer meetings. Liliana Danila, Felicia Emery and Maria Fitzpatrick provided excellent research assistance. Financial support was provided by the University of Virginia Bankard Fund for Political Economy and Clemson University College of Business and Behavioral Science.

1. Introduction

The 1990s saw the most dramatic reforms to U.S. welfare policy of the past half century, culminating in fundamental national reform in August 1996, when the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) replaced Aid to Families with Dependent Children (AFDC) with Temporary Assistance for Needy Families (TANF). The reforms aimed at promoting employment among recipients and reducing dependence on public assistance through time limits, job subsidies, work requirements, and increased funding for child support. While studies of welfare reform's short-term impacts on income and employment abound, little is known about its long-term effects.

This paper addresses one of the most important long-term impacts of welfare reform: its effect on the academic performance of children in low-income families. Successful reform will not only improve the economic well-being of disadvantaged families, but will also help sever intergenerational links in welfare dependency (Pepper 2000) by improving opportunities for children. Numerous studies have demonstrated that more education leads to higher earnings and better health.¹ Therefore, a key criterion by which to assess welfare reform is the educational performance of low-income children.

This study is the first attempt to measure the effects of national and state-wide welfare reforms on the academic performance of low-income children. We employ a state panel of math achievement levels from the National Assessment of Educational Progress (NAEP), covering the years 1996, 2000, 2003 and 2005, to estimate the reduced-form net effect of welfare reform. The task is inherently difficult: welfare reform, which occurred over an extended period of time including numerous economic and policy changes, does not provide an ideal experiment. We pursue several strategies designed to isolate the impact of welfare reform on low-income children. First, we use a difference-in-differences framework to estimate the change in academic performance for students in low-income families, with 1996 as the base year and more affluent students as a control group. We include a rich set of fixed effects to eliminate impacts of other policies that lead to trends in student performance nation-wide and state-by-state. We control for major categories of time-varying educational inputs, such as school spending and parental education, allowing for differential effects on low-income students. Finally, we address confounding effects due to contemporaneous

¹ Becker 1993, Zhang (forthcoming), Lleras-Muney 2004, among others.

policy changes that may affect low- and high-income students differentially, such as the expansion of the EITC and adoption of school accountability programs.

We find no evidence that welfare reform harmed the academic performance of students in low-income families. On the contrary, math scores of low-income 4th graders grew more rapidly in the post-reform era than scores of comparable high-income 4th graders. The estimated effects are statistically and economically significant for 2003 and 2005, and are largely insensitive to the inclusion or removal of covariates and fixed effects. The causal interpretation is strengthened when we exploit cross-state variation in the timing of reform and in its intensity. Larger effects occur in states with larger initial welfare caseloads and in states that experienced larger caseload reductions. Preliminary analysis associates these beneficial effects with relative improvements in low-income children's time use and interaction with parents. We find little variation in treatment effects by student ability, sex, or race. In a value-added framework, welfare reform is also associated with improved relative test score growth between 4th and 8th grades.

The existing welfare reform evaluation literature has focused primarily on outcomes for adults and households (Blank 2002, Moffitt 2003, Grogger and Karoly 2005). In the years following state and national welfare reforms, there is a sharp decline in welfare caseloads and a large increase in the labor supply of single mothers with children under 18 (e.g., Blank 2001, Hoynes 2000, Ziliak et al. 2000). There is some controversy regarding the effects of welfare reform on family income. While studies find evidence of rising incomes for female-headed households, others document income declines for low-education families (Bennett et al. 2004) and a rise in "deep poverty" (Haskins 2001). Meyer and Sullivan (2004) meanwhile show that consumption did not decline for vulnerable women as a result of welfare and tax reforms. Evidence on family structure is not conclusive and the effects have not been demonstrated with national data (Horn and Sawhill 2001, Murray 2001).

Long-run effects of welfare reform on low-income children's outcome have largely been ignored by scholars. The most compelling previous studies employed an experimental framework with random assignment into treatment (welfare reform) and control (traditional

welfare) groups, using data from five welfare demonstrations.² They find evidence of improvement in self- and teacher-reported academic achievement and test scores for younger children in the treatment group (Duncan and Chase-Lansdale 2001, Morris et al. 2005). For older children and adolescents, however, there is evidence of reduced academic performance measured by self-reported dropout, expulsion and suspension rates, and maternal assessment of overall achievement (Gennetian et al. 2004). Although proper randomization eliminates self-selection bias, these experiments were limited geographically, tended to employ weaker work requirements than those in state-wide and national programs, and could not capture effects due to welfare participation. These and other limitations may compromise the external validity of the conclusions from the experiments. By contrast, this paper measures the effects of state-wide and national welfare reforms on a nationally representative group of children from low-income families, including both current and previous welfare participants. Using a similar empirical approach, Miller and Zhang (2008) extends the population and outcomes of interest of this paper and confirms the main result. The paper finds positive treatment effects of welfare reform on the educational attainment of adolescents in low-income families.

2. Empirical Framework

School performance is generally modeled as the outcome of student ability, school inputs, and family inputs cumulative to the time of performance measurement (Hanushek 2002):

$$P_{it} = f(F_i^{(t)}, S_i^{(t)}, A_i) + v_{it}, \quad (1)$$

where P_{it} is the performance of student i at time t ; $F_i^{(t)}$ is a vector of family inputs and student effort cumulative to time t ; $S_i^{(t)}$ is school inputs cumulative to time t ; A_i is the innate ability; and v_{it} is a stochastic error term. Family inputs take effect once a child is born, and school inputs directly affect school performance while a child is in school.

Welfare reform can affect the school performance of low-income students through family inputs and student effort. Increased maternal employment can hinder child cognitive development by reducing time available for home production, such as supervising and disciplining children, reading to them, and assisting with homework. On the other hand,

² These were the Florida Family Transition Program; the National Evaluation of Welfare to Work Strategies in Atlanta (GA), Grand Rapids (MI) and Riverside (CA); the Minnesota Family Investment Program; the Milwaukee New Hope; and the Canadian Self-Sufficiency Project.

working mothers may feel more secure and confident, resulting in greater productivity at home. Working mothers may also provide their children with improved stability and daily routine, serve as positive role models for their children, and instill a desire for financial independence and greater academic achievement. Increased family income may also boost children’s school performance through improved nutrition and reading materials at home.³ Welfare reform may directly affect children’s learning effort, most importantly for adolescent girls, if they respond to the reduced availability and attractiveness of welfare as a long-term substitute for paid employment. These different channels may operate in different directions, and the net effect of welfare reform on school performance is inherently an empirical question.

This paper focuses on the net effect of welfare reform in a reduced-form analysis. We estimate a linear specification of the relation in Equation 1:

$$P_{ist} = \beta_s + \beta_{SLE} \cdot SLE_i + \beta_t \cdot \tau_t + \beta_{SLE,t} \cdot SLE_i \cdot \tau_t + \beta_X \cdot X_{ist} + v_{ist}, \quad (2)$$

where s indexes state, t indexes time, i indexes student race-sex-income group. We define the low-income treatment group as students eligible for federal (USDA) free or reduced-price lunches (SLE) and the higher-income control group as those ineligible for subsidized lunches (SLI). $SLE_i = 1$ if i is the SLE group. $\tau_t = 1$ if t is a post-reform year. X_{ist} includes proxies for a state’s school inputs, measured as averages for the period a group is in school, and proxies for a state’s overall trend in family inputs unrelated to welfare reforms, measured as averages between the birth year and the test year.

Our parameter of primary interest is the year-eligibility interaction term, $\beta_{SLE,t}$, an estimate of the net treatment effect on the SLE group from the year of welfare reform to year t . It is identified by within group variation over time, and captures the post-reform deviation of the low-income group’s performance from that of the high-income group, after controlling for the mean changes in the performance observed over that period. Test scores of higher-income students are used to impute counterfactuals for what would have happened to test scores of low-income students absent welfare reform. Because the income cutoff for subsidized lunches is 185% of the poverty level, the SLE group is a superset of the welfare

³ Prior studies of maternal employment and child human capital produce mixed evidence. Negative effects are concentrated among children in more affluent, two-parent families, with more educated mothers (Ruhm 2004, Blau and Grossberg 1992, Baum 2003, Gregg et al. 2005). For financial resources, Blau (1999) finds only trivial direct effects of family income on child development.

eligible, containing formerly and currently eligible children, as well as children from somewhat wealthier families. If the SLE proxy for the welfare eligible is overly inclusive or there exist positive cross-income peer effects, the $\beta_{SLE,t}$ estimate will be attenuated.

$B_{SLE,t}$ may increase with t for three reasons. First, the cumulative nature of Equation 1 implies that the effect of welfare reform will increase with duration of exposure. Since test scores measure the stock of knowledge accumulated through past investment, more years of exposure should produce larger effects. Second, there may be a lag between welfare reform and the full response in family inputs; for example, time limits take several years to become binding. Third, experimental studies find the greatest benefits for younger children, implying earlier exposure is beneficial. For 4th graders observed in the first decade following welfare reform, their age at first exposure decreases as time passes.

The empirical work addresses two potential sources of bias in the difference-in-differences (DD) analysis. First, other time-varying factors may cause differential changes in test scores for low- and high-income groups. Given the cumulative nature of the education production process, these confounding factors may be events that occurred before or after PRWORA. We identify two important policy changes: (1) the state-level school accountability reforms starting in the early 1990s and the passage of the No Child Left Behind (NCLB) legislation in 2002, and (2) the expansions of Earned Income Tax Credit (EITC) during the late 1980s and the early 1990s. In Sections 4-6, we present evidence that the treatment effects remain after controlling for these policy changes. Second, the compositions of the low- and high-income groups may not be stable over the period studied, mainly due to changes in the reporting of income status in the NAEP. We discuss this issue in the Appendix and present robust lower-bound estimates of the treatment effects that confirm the positive treatment effects of welfare reform.

3. Data

3.1 School Performance Measure

We measure school performance with mathematics test scores for 4th graders in the National Assessment of Educational Progress (NAEP) program.⁴ The 500-point test is given to a representative sample of about 3,000 4th graders in each state between late January and early March of the test year. It provides an independent measure of knowledge and aptitude of U.S. students in mathematics and has been widely used for cross-state and time-series comparison of student performance. We use math test scores for 1996, 2000, 2003, and 2005, the only years NAEP reports state average test scores separately for the SLE and SLI groups. Each observation is a state-year-race-gender-lunch eligibility cell.

Over the ten-year period between 1996 and 2005, there has been an across-the-board improvement in 4th graders' math test scores. The pattern of test score gains, however, differs substantially between the SLE and SLI groups, as illustrated in Figure 1, which plots the average test score gains for each group relative to its 1996 average. The gains for the SLE group are larger than those for the SLI group for the entire period, and the difference is largest for 2003 and 2005. Thus, on top of the overall trend of increasing test score, the SLE group experiences an even larger increase, which, given the timing of the welfare reform, may be attributable to the impact of welfare reform on family inputs for education production. The analysis to follow aims at establishing firm evidence for a causal interpretation. The increasing gains of the SLE group over time also suggest a lag between welfare reform implementation and its impact on low-income children's test scores. Detailed summary statistics are reported in Appendix Table 1.

3.2 Welfare Reform Variables

We exploit both cross-sectional and time-series variations in state welfare policies. States implemented state-wide AFDC waivers and TANF at different dates since the early 1990s; they also adopted a myriad of policy changes under the waivers and TANF, which also evolved over time. Therefore, low-income families in different states are exposed to welfare reform, and to various specific policies, for different lengths of time.

⁴ Test scores are shown to have a significant effect on wages even after controlling for education attainment (Murnane et al. 1995, Hanushek and Zhang 2006). Math test scores are also strong predictors of educational attainment (Murnane et al. 1995).

Data on welfare policies are from Crouse (1999), U.S. DHHS (1997) and the Urban Institute Welfare Rules Database. The first two columns of Appendix Table 2 show the earliest implementation dates for state-wide AFDC waivers and TANF. Eleven states implemented state-wide waivers before 1995. The next six columns show the adoption and evolution of the sanction policy, penalties on welfare recipients for violating work requirements. We focus on this specific policy because it is a well-defined and theoretically important dimension that can potentially provide a direct and immediate incentive for welfare recipients to engage in labor market activities. Additionally, it has sufficient variation over time.⁵ We define four types of sanctions based on the reduction in cash benefits the first time work requirements are violated (initial sanction) and on the most severe sanction for violation. “No sanction” was the default before AFDC waivers. States started to implement sanction policies under either their AFDC waiver or TANF. Twenty-nine states initially adopted a “partial sanction” (benefits are reduced partially in both the initial and the most severe sanctions); half the remaining had a “gradual sanction” (benefits are reduced partially in the initial sanction and fully in the most severe sanction), and the other half a “full sanction” (benefits are reduced fully in both the initial and the most severe sanctions). Over time, 22 states switched to more stringent sanctions, and no state reverted to a less stringent one.

3.3 Time-Varying State Characteristics

We include a vector of state characteristics to account for the major categories of time-varying educational inputs. Pupil-teacher ratio and expenditure per student are proxies for measurable school inputs; average income, unemployment rate, percentage of population 25 years of age or older with a high-school diploma or higher and percentage with a bachelor’s degree or higher reflect forces that may affect family inputs. School input variables are averages over the years a typical 4th grader is in school: 1992-1995, 1996-1999, 1999-2002, and 2001-2004 for test years 1996, 2000, 2003, and 2005 respectively; family input variables are averages between the birth year of a typical 4th grader and the test year: 1987-1995, 1991-1999, 1994-2002, and 1996-2004. Appendix Table 3 reports summary statistics of these

⁵ Isolating the effect of different features of welfare policies is a well-known challenge (Blank 2002). Not only are there a large number of policy dimensions, but state policy choices also tend to be highly correlated. Therefore, one should be cautious in interpreting the findings about sanction policies in Section 5.

variables for each test year. During this 10-year period, there is an overall increase in school inputs reflected by increasing expenditures per student and decreasing pupil-teacher ratios. State income and education levels have also increased overall.

3.4 Introduction of State School Accountability Systems

States began introducing school accountability systems in the early 1990s, and their effect on academic performance is unlikely captured by measurable school inputs. The uneven introduction of accountability systems over time, and difference in timing of the two types of reforms, allows us to separately identify the effects of accountability systems and welfare reform on school performance.

Following Hanushek and Raymond (2005), an accountability system is defined as a mechanism for publicly disseminating information on standardized test performance for each school, along with a way to aggregate and interpret the school performance measure. States are classified as “consequential” states if they both report results and attach consequences to school performance or “report card” states if they only provide a public report. Consequential accountability may provide stronger incentives to schools than report-card accountability. The last two columns of Appendix Table 2 report the year each state introduced either type of accountability system, ranging from 1993 to 2003. By 2003, 31 states had consequential accountability.

The passage of No Child Left Behind Act (NCLB) in January of 2002 demanded stronger accountability of schools in all states. Between January and June 2003, all states had submitted to the Department of Education the required state plan for implementing its accountability system under NCLB, which were all approved by June 2003. Therefore, we consider 2003 the year the 19 report-card states and DC introduced consequential accountability. We define a dummy variable equal to 1 if an accountability system was in place prior to a NAEP test year; we also create a variable for the years elapsed since a state implemented an accountability system.

4. Effects of Welfare Reform on Fourth Grade Math Scores

This section compares the variation in test scores of low- and high-income children before and after the 1996 national welfare reform legislation to establish the main results. Included

in the analysis are states that did not implement any state-wide waiver before 1995 (non-waiver states). If there is a lag between implementation of a waiver policy and its impact on children's academic performance, the 1996 test scores in non-waiver states are unlikely to be affected by waivers.⁶

4.1 Average Effects of National Reform

We first estimate Equation 2 using state average 4th grade math score as the performance measure. Results in Table 1 show a consistently positive effect of welfare reform on low-income 4th grader's math scores in the decade following national welfare reform. All specifications include state fixed effects. All coefficients are reported with robust standard errors, clustered at the state level, allowing for arbitrary error term heteroskedasticity and correlation within a state.

The specification in Column 1 includes only fixed effects. The positive interaction terms, indicating larger gains for free-lunch eligible students relative to ineligible students, are increasing over time and significantly different from zero (at the 1% level) for 2003 and 2005. The other coefficients indicate that males outperformed females by 1.4 points, whites outperformed non-whites by 15 points, and SLE students lagged SLI students by 18 points. The significant year effects show that 4th grade math scores were improving overall during the period.

Column 2 adds key time-varying state education inputs: family inputs measured by adult population shares with high school diplomas and college degrees, per capita income (in tens of thousands of 1983 dollars) and unemployment rate; school inputs measured by expenditures per pupil (in tens of thousands of 1983 dollars) and pupil-teacher ratio.⁷ In addition, we address the concern that school accountability reform and NCLB tend to focus on low achievers while aiming at improving all performance. We include an indicator for the presence of consequential accountability programs, a linear measure of years since the

⁶ Using 1995 as the cutoff year is somewhat arbitrary; the results, however, are robust to using 1994 or 1996 as the cutoff year. Analysis in this and the remaining sections excludes DC. DC is heavily concentrated with low-income and minority students, and has long stood out on the NAEP tests for its poor performance. Including DC increases the magnitude and significance of the treatment effect estimates.

⁷ Additional controls for Medicaid coverage rates of children 10 and under are obtained from March CPS files, averaged over the relevant period and interacted with SLE status. These variables do not independently influence test scores or meaningfully affect the magnitude or significance of the estimated treatment effects.

accountability program was initiated, and interactions of SLE with each of these.⁸ Accountability has a positive and significant effect on test scores, and years since accountability has a positive but insignificant effect. The magnitude of estimates is comparable to Hanushek and Raymond (2005). The SLE-accountability interactions are both negative and insignificant. The estimates for common terms in the first two columns are quite similar, and welfare reform is again shown to have a positive effect on relative test scores, which is statistically significant starting in 2003, and somewhat larger in magnitude.

Column 3 adds a complete set of fixed effects for state and year interactions to remove any state specific time trends in test scores. The comparison is the performance gap between SLE and SLI students within a state and a year, and the $\beta_{SLE,t}$ terms capture average changes in these gaps from 1996 to 2000, 2003 and 2005. Column 4 includes further interactions between time-varying state education inputs and SLE indicator, thereby allowing them to have different effects on low- and higher-income students. None of the interactions is individually significant. In both specifications, our main variables of interest, the $\beta_{SLE,t}$ terms, are again positive, increasing, and significant starting in 2003.⁹

Column 5 represents a preliminary analysis to further disentangle the effects of NCLB and welfare reform for 2005, the first NAEP test year after the implementation of NCLB. NCLB requires states to disaggregate test results by socio-economic group, while states may or may not choose to do so under their school accountability systems. Therefore, NCLB may have a stronger differential effect on SLE and SLI students than state accountability, and designating all states as consequential states in 2005 may not capture this extra impact.

To isolate NCLB's extra differential effect, we would ideally like to control for the differential pressure of NCLB on SLE and SLI students or their schools. A direct measure being unavailable, we use as a proxy the share of 4th graders whose math scores in 2003 state tests fell into one of the achievement categories adjacent to the proficiency cutoff, proficient

⁸ Report-card programs do not have a significant effect on performance, either alone or added along with consequential programs; nor do they affect estimated effects of welfare reform.

⁹ The NAEP state sample is representative of 4th graders in each state; therefore the underlying cell size of each race-gender-income observation varies within and across states. To address the potential heteroskedasticity and improve the efficiency of our estimates, we repeat the regressions using as a weight the percentage of each race-gender-income group in state 4th-grade population. Furthermore, to produce estimates that are representative of the average student in the nation, we use as a weight the percentage of each race-gender-income group in the U.S. 4th-grade population. The estimated effects of welfare reform are unchanged.

or basic, for each income category and each state.¹⁰ The hypothesis is that NCLB pressure leads schools and teachers to focus resources on students near the cutoff, rather than those far below (scoring below basic) or far above (scoring advanced) it. These shares vary across states for two reasons: (1) differences in achievement levels of students and (2) differences in the stringency of state tests and standards. The interaction of share-near-cutoff with year 2005 indicator is added in the regression; as shown in Column 5, it does not have a significant effect on test scores, and the estimates of $\beta_{SLE,t}$ terms remain almost identical to those in Column 4.

We conducted other robustness checks. First, based on different behavioral assumptions about schools and teachers, we use shares of students below basic, basic, and both below basic and basic to measure pressures produced by NCLB for 2005. None of these has an impact on test score or affects the size and significance of the treatment effect estimates. Second, concerned that these NCLB pressure measures reflect more student ability than stringency of state standard,¹¹ we use the state proficiency standard stringency measure created by NCES (2007), either directly as a control variable or as an instrumental variable for our NCLB pressure measure. This reduces our sample size almost by half; yet again, the stringency measure has no effect on test scores, and the treatment effects remain unchanged. Although due to data limitation we could not establish a firm relationship between NCLB and 2005 test scores, the robustness of the treatment effect estimates provides support for the validity of the natural experiment framework.

Welfare reform has a nontrivial effect on test scores, irrespective of model specification. Consider the estimates in Column 2 of Table 1. Relative to the baseline trend between 1996 and 2005, the SLE students experienced 25% greater test score growth. This extra growth is equivalent to 0.096 of a standard deviation of the initial 1996 test score distribution of the SLE group.¹² To interpret the magnitude of gains from welfare reform, it is useful to compare them to gains from changes in school inputs. Krueger (1999), studying the Tennessee STAR program, finds that a one third (8 students per class) reduction in class size increases students' test score by 0.17 of a standard deviation in kindergarten year.

¹⁰ The data are obtained from each state Department of Education webpage. For a few states without 2003 information, we use the share for 2004.

¹¹ All of the share measures, however, have low and insignificant correlations with the 2003 NAEP scores.

¹² The standard deviation of the initial 1996 test score distribution of the SLE students from public schools nation-wide is 31.

Rivkin, Hanushek, and Kain (2005) find somewhat smaller gains for 4th through 6th graders in Texas public schools. In both studies, the effect on free-lunch eligible students is at most slightly larger than that for ineligible students. Thus, welfare reform's effect on low-income students appears to be significant quantitatively as well as qualitatively. Furthermore, these estimates may understate the effects on program participants, since welfare eligible children account for fewer than half of all SLE students. The initial 1996 test score gap between SLE and SLI groups was narrowed by 17% in 2005. From a policy perspective, it is essential to note that welfare reform reduces public expenditures, while the comparison programs require increased spending.

For most of the analysis, we employ an aggregate unit of observation, corresponding to the published NAEP values and consistent with the state-year control variables. However, in order to estimate treatment effects for a more intense treatment group, namely children who qualify for free lunches (FLE, family income below 130% of the poverty line, rather than 185% for SLE), we obtained confidential individual micro-data from the NAEP. We re-estimated the full model using the micro-data, with separate fixed effects and interactions for free-lunch and reduced-price lunch eligible children. The estimated treatment effects for the free-lunch group and key covariates are reported in Column 6 of the table, with robust standard errors clustered at the state level. As expected, the post-welfare interaction terms are larger for the free-lunch treatment group than for the subsidized and free-lunch group in the previous columns. The treatment effect is now statistically significant by the first observed year following welfare reform, 2000. The baseline test score gap between low- and higher-income children increases substantially (from 16 to 36 points), however, and welfare reform only narrows the gap by 11%.

4.2 Heterogeneous Effects

Previous estimates using cell average score as the dependent variable may conceal important variation among students of different abilities or suffer from attenuation bias if there is substantial clumping of scores at the extremes of distribution. More importantly, since SLE students have lower initial test scores, the positive effects may be the outcome of mean reversion.

In this section, we present separate estimates for students at 4 points on the test score distribution. The treatment and control groups in each case are designed to be comparable in their achievement before welfare reform. Specifically, we match students at the 25th SLE percentile with those at the 10th SLI percentile, the median SLE with the 25th SLI, the 75th SLE with the median SLI, and the 90th SLE with the 75th SLI. As shown in the top panel of Table 2, SLE and SLI groups in each of the four samples thus created have similar initial 1996 test scores. Estimation results from the four samples are reported in Table 2. As above, welfare reform is associated with a relative improvement in test scores for low-income children, generally statistically significant in 2003 and 2005, and somewhat larger for children at the lower end of the test score distribution. These findings confirm that beneficial effects of welfare reform on low-income children are neither due to mean reversion nor driven by uncontrolled factors that operate through initial test score differences between the groups.

We also estimate welfare reform's effect separately for boys and girls, Whites and Non-Whites. Welfare reform may have a more positive effect on the test scores of girls than boys if girls are more influenced by working mother role models or if they respond more to the diminished attractiveness and feasibility of welfare as a permanent substitute for employment. Alternatively, boys may experience greater gains if welfare reform changed family structure in such a way that boys became more likely to have contact with their fathers. Results for two specifications by sex are reported in Panel A of Table 3, corresponding to the basic model (Column 2 in Table 1) and the full model (Column 4 in Table 1). The estimates are generally larger for girls, but not statistically significant so. Results by race, reported in Panel B, show no consistent evidence of heterogeneous treatment effects. In the basic specification, the treatment effects are larger for Whites, while they are larger for Non-Whites in the full specification. The rest of the paper will focus on overall effects.

4.3 Preliminary Analysis of Changes in Student Behavior and Home Inputs

To explore mechanisms generating the relative test score gains following welfare reform, we examine changes in the self-reported behavior and home environments of 4th grade NAEP test-takers between 1996 and later years for both SLE and SLI groups

In response to the question “How often do you discuss things you have studied in school with someone at home?” students in both SLE and SLI groups report higher average values in 2005 than in 1996, with a far larger increase for SLE than SLI students. The mean frequency goes from 8.2 to 10.0 discussions per month for SLE, a 22% increase, and from 9.1 to 10.2 for SLI, an 11% increase.¹³ The comparison is imperfect due to a change in the multiple-choice response options starting in 2003; however, the “never or hardly ever” and “1-2 times a month” categories were offered in all years, and we therefore compare changes at the low end of the distribution. Both groups exhibit increasing shares of students in the lowest category, but increased representation in the “zero” bin is 4 times as large for SLI as for SLE when calculated in percentage point terms (4% to 1%) and 6 times as large when measured as a percent change from 1996 values (31% to 5%). Hence, the movement at the left tail corroborates the movement of the mean. When we estimate the effect in our micro-level difference-in-differences framework, with the full set of controls, the treatment effects (FLE*year interactions) are positive, and statistically significant starting in 2003. The linear model estimates a relative increase of about 2 conversations per week. Ordered Probit estimates were comparable. We lack information regarding the content or duration of these discussions; yet, this measure at the extensive margin suggests that welfare reform may be associated with greater parental investment in child education in low-income households.

Students also reported the amount of time watching television and video on school days. This information sheds light on changes in time inputs for building human capital. In 1996, over a quarter of SLE students and more than one eighth of SLI students report the highest option of “6 hours or more.” Between 1996 and 2003, the average number of hours drops from 3.86 to 3.27 for SLE children (a decline of 35.4 minutes per day, or 15%), and from 3.15 to 2.91 for SLI children (a 14.7 minute, or 7%, decline).¹⁴ The full micro-level difference-in-differences framework again yielded similar results, with a statistically significant relative decline of about 15 minutes for FLE children by 2003. This relative decrease in television viewing by FLE students can imply a relative increase in their time spent on school-work and related activities. This may be the result of improved parental

¹³ To compute the means, we assign a numerical value for frequency per 4 week period to each categorical response option: never or hardly ever = 0; 1-2 times a month = 1.5; 1 time a week = 4; 1-2 times a week = 6; 2-3 times a week = 10; almost every day = 13; every day = 24.

¹⁴ Means are computed by assigning numerical values to categorical response options: none = 0; 1 hour or less = 1; 2 hours = 2; 2-3 hours = 2.5; 3 hours = 3; 4 hours = 4; 4-5 hours = 4.5; 5 hours = 5; 6 hours or more = 7.

disciplining and role modeling or enhanced stability and daily routine in low-income households following welfare reform.

5. Robustness: Test Score Gains Linked to Welfare Reform

Section 4 exploits variation in test scores before and after *national* welfare reform to establish the main results. This section exploits *cross-state* variation in welfare reform timing and intensity to provide additional identification, and uses a longer time period to test for pre-existing trends.

5.1 Test Score Trends Following Welfare Reform

This section establishes the robustness of the treatment effects by exploiting cross-state variation in the timing of implementing state-wide AFDC waivers and TANF policies in general, and the variation in both timing and severity of the sanction policy under AFDC and TANF in particular for all states. This approach also further disentangles the effects of welfare reform from school reforms, given that the early waiver states are quite different from the early accountability states as shown in Appendix Table 2.

We estimate the differential relationship between test scores and time elapsed since any state-wide reform for low- and higher-income students as modeled in Equation 3:

$$P_{ist} = \beta_s + \beta_{SLE} \cdot SLE_i + \beta_t \cdot \tau_t + \beta_{SLEMSR} \cdot SLE_i \cdot MSR_{st} + \beta_{MSR} \cdot MSR_{st} + \beta_X \cdot X_{ist} + v_{ist}, \quad (3)$$

where our variable of interest is $SLE_i \cdot MSR_{st}$, the interaction term between MSR_{st} (months since reform) and free lunch eligibility, and β_{SLEMSR} reflects test score growth of SLE students relative to SLI students for each month that passed following welfare reform. Although the impact of welfare reform may not be linear, the limited samples preclude any investigation of nonlinearities or the simultaneous inclusion of the SLE*year interaction terms.

The treatment effect estimates reported in Table 4 are invariably positive and statistically significant across specifications. The first 4 columns have the same specifications as the first 4 columns in Table 1: Column 1 includes only fixed effects; Column 2 adds time-varying state education inputs and variables for school accountability reforms; Column 3 adds state-year interaction terms; and Column 4 adds interactions between time-varying education inputs and SLE indicator. For each month that passed following welfare reform, SLE students saw an extra growth in test score of between 0.015 and 0.02 point relative to SLI

students. The size of the treatment effect is quite close to the monthly rate implied by the main result of last section: dividing the total effect between 1996 and 2005 ($\beta_{SLE,2005}$ in Table 1) by the number of months passed since 1996 (108) produces a range of estimates from 0.023 to 0.029.

Other estimates are comparable with the results in Table 1 for the non-waiver states. Whites score higher than Non-Whites, boys score higher than girls, and SLE students score lower than SLI students. There is a positive (significant) and growing (insignificant) benefit from accountability on math scores. Estimates of β_{MSR} are statistically insignificant, confirming that welfare reform has no effect on SLI students.

The final column is an attempt to address a new concern that arises for this method – the potential endogeneity of the timing of state reforms. The estimates of β_{SLEMSR} in the previous columns are identified from variation both within states over time, and across states due to their timing of reform implementation. If there is a correlation between timing of implementation and the initial test score gap between SLE and SLI students, β_{SLEMSR} will be biased. In particular, if states with smaller test score gaps implemented reforms earlier, β_{SLEMSR} will be biased upwards. In column 5, we address this concern by adding interactions between SLE and state indicators. These terms remove any average difference in test score gaps across states. The new estimate of β_{SLEMSR} is 0.047 with a standard error 0.015, showing that the effect is robust to controlling for differences in levels.

We investigate the effect of sanction policy timing and severity in Equation 3, by replacing MSR_{st} and $SLE_i \cdot MSR_{st}$ with a new set of policy variables. From the information in Appendix Table 2, we construct the variables $MSS1_{st}$ (months since any sanctions), $MSS2_{st}$ (months since gradual or full sanctions), and $MSS3_{st}$ (months since full sanctions), which are interacted with SLE_i to produce the new variables of interest, in increasing order of severity: $SLE_i \cdot MSS1_{st}$, $SLE_i \cdot MSS2_{st}$, $SLE_i \cdot MSS3_{st}$. The coefficients, $\beta_{SLEMSS1}$, $\beta_{SLEMSS2}$, and $\beta_{SLEMSS3}$, have similar interpretation as β_{SLEMSR} but allow heterogeneous effects by sanction severity. Estimates are presented in Table 5, which mirrors the structure of Table 4.

The estimates for $\beta_{SLEMSS1}$ are largely positive and significant at the 10% level across specifications, showing that even partial sanctions produced a positive effect on test scores. The estimates for $\beta_{SLEMSS2}$ and $\beta_{SLEMSS3}$ are positive but insignificant, and F-tests on $\beta_{SLEMSS2} + \beta_{SLEMSS3}$ fail to reject zero. This implies that states with more severe penalties for violations

of work requirements had test score gains that were statistically indistinguishable from those with partial sanctions. The lack of precision in the estimated effects of sanction severity may result from measurement error – we measure the rules, but actual enforcement may be more important. The noise surrounding these estimates underscores the difficulties of empirically isolating the effects of particular dimensions of welfare policy.

5.2 Treatment Effect Size and Welfare Caseload

The above estimates of the treatment effect are obtained by linking test score gains for free-lunch and subsidized-lunch eligible students to changes in welfare rules. The link between the treatment effect and welfare reform is weakened for two reasons. (1) Only a subset of SLE students is affected by welfare reform. (2) Seemingly similar policy rules may conceal large differences in the actual enforcement of policies and their effectiveness in altering the behavior of current and potential welfare recipients. Data limitations preclude a direct tackling of these problems. As an alternative, we provide supportive evidence to further link the treatment effect to welfare reform.

We focus on correlations between the treatment effect and the 1996 caseload ratio and the change in caseload ratio between 1996 and 2005 across states. The caseload ratio is defined in each state as the ratio of the number of families receiving cash welfare benefits to the number of children in that state receiving subsidized meals from the school lunch program. States with higher caseload ratios in 1996 have greater fractions of SLE students likely to be affected by changes in welfare rules. States with larger caseload declines between 1996 and 2005 are likely to have enforced their welfare policy changes more effectively. Each of the two variables measures an aspect of the treatment intensity experienced by the SLE students. The treatment effect is expected to be larger for states in which the SLE treatment group experienced more intense treatment from welfare reform.

The treatment effect is computed for each state using the $SLE*(Year = 2005)$ term in a regression model of 4th grade math scores on student characteristics and state-year varying controls, including accountability reform. Caseload ratios are residuals after adjusting for state-year varying controls. Using non-waiver states as observations, the correlation between the treatment effect and the adjusted 1996 caseload ratio is 0.31 and significant at the 7.6% level, and the correlation between the treatment effect and the change in adjusted caseload

ratio between 1996 and 2005 is -0.32 and significant at the 6.7% level. When we narrow the treatment group to include only FLE children, and re-define the state caseload ratio to the number of families receiving cash welfare benefits divided by the number of children receiving free lunches, the relationships are even stronger. The correlation between treatment and adjusted 1996 caseload ratio increases to 0.362 (significant at the 4% level) and the correlation between the treatment effect and the change in adjusted caseload ratio between 1996 and 2005 becomes more negative at -0.50 and significant at 0.3%. In sum, the observed cross-state correlations reinforce the positive association between welfare reform and relative math performance established earlier.

5.3 Pre-Existing Trends in the Achievement Gap

We now turn to another concern: the observed narrowing during our sample period may not be caused by welfare reform, but may instead be the result of a pre-existing trend of a narrowing achievement gap by income. Because test scores by free lunch eligibility are not available before 1996, we address this concern indirectly by comparing test score trends of different racial and ethnic groups. Eligibility for the federal lunch program is highly correlated with race: SLE students constitute 23% of Whites, 72% of African Americans and 69% of Latinos. Figure 2 provides suggestive evidence against a pre-existing narrowing by plotting the changes in national average test scores relative to the 1996 baseline for White, African American, and Latino students.¹⁵ In the years preceding welfare reform, test scores of white students improved relative to those of Latino students and were stable relative to those of African American students. However, after 1996, and especially after 2000, African American and Latino students showed relative test score gains. The time pattern is consistent with a break from trend and supports a causal interpretation for the findings of the paper. In a related paper, Miller and Zhang (2008) find relative gains in the academic attainment of low-income children in the years following welfare reform, but no evidence of relative gains in the years preceding welfare reform.

¹⁵ National data were used in order to include the year 1990. The values for test score gains relative to 1996 for non-waiver states in the years 1992, 1996, 2000, 2003, 2005 are as follows. Whites: -5, 0, 3, 12, 15. African Americans: -6, 0, 7, 17, 21. Latinos: -4, 0, 6, 16, 20. Eliminating the early waiver states clearly has the effect of diminishing the relative gains for African Americans between 1992 and 1996.

6. Value-Added Estimation

Finally, we are concerned that the treatment effect for SLE students estimated in the previous sections may in part be attributable to the 1993 and earlier expansions of the earned income tax credit (EITC) that also affected the income and work incentives of low-income families. EITC expansions altered the effective wages of low-earners, were generally found to increase the labor supply of low-income families (Hotz and Scholz 2003), and have been associated with improvements in children's academic achievement (Dahl and Lochner 2005). These effects are of the same direction as those of the welfare reform; not controlling for EITC will bias the estimated effect of welfare reform upward. The timing of EITC expansions implies that they will affect school performance in our sample mostly through their effects on family inputs during early childhood.

We estimate Equation 2 in a value-added framework to provide supportive evidence to our baseline estimation of the treatment effect. We exploit the fact that the 1987 cohort was tested in 1996 as 4th graders and in 2000 as 8th graders, and the 1991 cohort was tested in 2000 as 4th graders and could potentially be tested in 2004 as 8th graders.¹⁶ Therefore, if we estimate the performance of 8th graders in 2000 and 2004 while controlling for their respective performance in 1996 and 2000 as 4th graders, we can remove the effects of confounding factors that take place during early childhood and during the first four years of schooling. The idea is illustrated as follows:

$$P_{it}^{g=8} = f(P_{i,t-4}^{g=4}, F_i^{(t-4,t)}, S_i^{(t-4,t)}) + v_{it} - v_{i,t-4}, \quad (1')$$

where $t = 2000$ and 2004 for cohorts born in 1987 and 1991.¹⁷ Note that Equation 1' also eliminates anything that appears as a fixed effect in Equation 1. We carry out the estimation in a linear model similar to Equation 2:

$$P_{ist}^{g=8} = P_{is,t-4}^{g=4} \cdot \alpha + \beta_s + \beta_{SLE} \cdot SLE_i + \beta_t \cdot \tau_t + \beta_{SLE,t} \cdot SLE_i \cdot \tau_t + \beta_X \cdot X_{ist} + v_{ist} - v_{is,t-4}. \quad (4)$$

¹⁶ State average NAEP math test scores for 8th grade SLE and SLI students are available for 1996, 2000, 2003, and 2005. We estimate Equation 2 for 8th grade math. In the full model, the estimates for SLE*2000 and SLE*2003 are smaller than their counterparts in the 4th grade regression and are insignificant, but the estimate for 2005 is significant and of comparable magnitude to that in the 4th grade regression. The results for 4th and 8th graders jointly suggest that the beneficial effects of early exposure to the welfare reform do not disappear over time. Regression results from Equations 2 and 3 for 8th graders are reported in Appendix Table 4.

¹⁷ Since NAEP did not test the 8th graders in 2004 (1991 cohort), we use the average of the test scores of the 8th graders in 2003 and 2005 (1990 and 1992 cohorts) as observations for 2004.

In Equation 4, the relevant family inputs for the 1987 cohort and the 1991 cohort are averages between 1996 and 1999 and between 2000 and 2003 respectively. Therefore, for this approach to remove the effects of the EITC, we need to assume that the 1993 EITC expansion has the same effect on family inputs for both of the 4-year periods; for example, it has limited lagged effects on family inputs (lag < 3 years) or begins to affect family inputs within three years of its passage. If this assumption holds, $\beta_{SLE,2004}$ will measure welfare reform's effect accumulated between 2000 and 2004 relative to that accumulated between 1996 and 2000.

Partial estimation results of Equation 4 from our full specification are reported as follows:

$$P_{ist}^{g=8} = P_{is,t-4}^{g=4} \cdot 1.09 + 3.52 \cdot 1(SLE_t = 1) \cdot 1(t = 2004) + \dots + \hat{\varepsilon}_{ist} \quad (0.3) \quad (1.77)$$

Numbers in parentheses are robust standard errors clustered at state level. The control for 4th-grade score removes differences in initial stock of human capital, and its coefficient is close to unity. Relative to welfare reform's effect on the growth of math scores between 1996 and 2000, its effect between 2000 and 2004 is 3.52 points and significant at the 10% level. This suggest that the EITC expansions cannot fully explain the narrowing of the test score gap between SLE and SLI students, and that welfare reform played a role. Coefficient estimates on other control variables are generally consistent with those in Table 1. When we examine male and female, White and Non-White separately, the point estimates range from 3.22 to 3.99.

Although this exercise provides additional support for the causal interpretation of welfare reform's effect on test score gains of low-income students, data limitations prevent a more conclusive strategy for separating the effects of EITC expansions. Regardless, the basic structural implications remain unchanged even if the EITC contributed to the measured test score gains following welfare reform. The two policies may affect the labor supply of different demographic groups within the low-income population and have different treatment intensities; yet, the results from either policy imply that increased parental work hours help students in low-income families.

7. Conclusion

This paper presents the first analysis of the impact of the 1996 federal welfare reform on the academic achievement of children in low-income families using nationally representative data that spans a ten year period starting just prior to the welfare reform. We estimate the net effect of welfare reform in a reduced form analysis. In a difference-in-differences framework, we use children in higher-income families as a control group to simulate the unobservable counterfactual time trend, and control for major time-varying educational inputs. There is no evidence that welfare reform harmed the academic performance of low-income students. On the contrary, the key finding is an improvement in the relative math scores of fourth-grade students from low-income families that is present for all race, gender and ability groups.

Preliminary analysis suggests that these beneficial effects are associated with relative improvements in low-income children's time use and interaction with parents. These findings seem to suggest that the harm due to less maternal time available to invest in children's human capital caused by increased employment is more than offset by gains in the quality of parental inputs and changes in child inputs. However, the reduced-form approach in this paper is unable to pinpoint a single underlying mechanism responsible for the beneficial effects. Welfare reform led to multiple behavioral changes in low-income parents, including increased labor force participation and reduced welfare receipt. A thorough exploration of the channels for the test score gains is a promising avenue for future research.

References

- Baum, C.L., II, *Does Early Maternal Employment Harm Child Development? An Analysis of the Potential Benefits of Leave Taking*. *Journal of Labor Economics*, 2003. 21(2): p. 409-48.
- Becker, G.S., *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. 3rd Edition, the University of Chicago Press, 1993.
- Bennett, N.G., H.-H. Lu, and Y. Song, *Welfare Reform and Changes in the Economic Well-Being of Children*. *Population Research and Policy Review*, 2004. 23(5-6): p. 671-99.
- Blank, R.M., *Declining Caseloads/Increased Work: What Can We Conclude about the Effects of Welfare Reform?* Federal Reserve Bank of New York Economic Policy Review, 2001. 7(2): p. 25-36.
- Blank, R.M., *Evaluating Welfare Reform in the United States*. *Journal of Economic Literature*, 2002. 40(4): p. 1105-66.
- Blau, D.M., *The Effect of Income on Child Development*. *Review of Economics and Statistics*, 1999. 81(2): p. 261-76.
- Blau, F.D. and A.J. Grossberg, *Maternal Labor Supply and Children's Cognitive Development*. *Review of Economics and Statistics*, 1992. 74(3): p. 474-81.
- Crouse, G., *State Implementation of Major Changes to Welfare Policies, 1992-1998*, 1999. Available at http://aspe.hhs.gov/hsp/Waiver-Policies99/policy_CEA.htm.
- Dahl, G. and L. Lochner, *The Impact of Family Income on Child Achievement*. NBER working paper #11279, 2005.
- Duncan, G.J. and P.L. Chase-Lansdale, *Welfare Reform and Children's Well-Being*, in *The new world of welfare*, R.M. Blank and R. Haskins, Editors. 2001, Brookings Institution Press: Washington, D.C. p. 391-412.
- Fletcher, S.H. and M.E. Raymond, *The future of California's academic performance index*. CREDO, Hoover Institution, Stanford University, 2002.
- Goertz, M.E. and M.C. Duffy, *Assessment and Accountability Systems in the 50 States: 1999-2000*. RR-046, Consortium for Policy Research in Education, Graduate School of Education, University of Pennsylvania, 2001.

- Gregg, P., E. Washbrook, C. Propper, and S. Burgess, *The Effects of Mother's Return to Work Decision on Child Development in the UK*. *Economic Journal*, 2005. 115(501): p. F48-F80.
- Grogger, J. and L.A. Karoly, *Welfare Reform: Effect of a Decade of Change*. Harvard University Press, 2005.
- Gennetian, L., G. Duncan, V. Knox, W. Vargas, E. Clark-Kauffman, and A. London, *How Welfare Policies Can Affect Adolescents: A Synthesis of Evidence from Experimental Studies*, *Journal of Research on Adolescence*, 2004. 14(4): p. 399-423.
- Hanushek, E.A., *Publicly Provided Education*. in *Handbook of Public Economics*, Alan J. Auerbach and M. Feldstein, editors. Amsterdam: North-Holland, 2002.
- Hanushek, E.A. and M.E. Raymond, *Does School Accountability Lead to Improved Student Performance?* *Journal of Policy Analysis and Management*, 2005. 24(2): p. 297-327.
- Hanushek, E.A. and L. Zhang, *Quality-Consistent Estimates of International Returns to Skill*. National Bureau of Economic Research, Inc. 2006, NBER Working Papers: w12664.
- Haskins, R., *Effects of Welfare Reform on Family Income and Poverty*, in *The new world of welfare*, R.M. Blank and R.E. Haskins, Editors. 2001.
- Horn, W.F. and I.V. Sawhill, *Fathers, Marriage, and Welfare Reform*, in *The new world of welfare*, R.M. Blank and R.e. Haskins, Editors. 2001.
- Hotz, V.J. and J.K. Scholz, *The Earned Income Tax Credit*. In Robert A. Moffit ed *Means-Tested Transfer Programs in the United States*, the University of Chicago Press, 2003.
- Hoynes, H.W., *Local Labor Markets and Welfare Spells: Do Demand Conditions Matter?* *Review of Economics and Statistics*, 2000. 82(3): p. 351-68.
- Krueger, A.B., *Experimental Estimates of Education Production Functions*. *Quarterly Journal of Economics*, 1999. 114(2): p. 497-532.
- Lleras-Muney, A., *The Relationship between Education and Adult Mortality in the United States*. *Review of Economic Studies*, 2005. 72 (1): p. 189-221.
- Meyer, B.D. and J.X. Sullivan, *The Effects of Welfare and Tax Reform: The Material Well-Being of Single Mothers in the 1980s and 1990s*. *Journal of Public Economics*, 2004. 88(7-8): p. 1387-1420.
- Miller, A. R. and L. Zhang, *Intergenerational Effects of Welfare Reform*, Mimeo, 2008.

- Moffitt, R., *Economic Effects of Means-Tested Transfers in the U.S.*. National Bureau of Economic Research, Inc. 2002, NBER Working Papers: w8730.
- Moffitt, R.A., *The Temporary Assistance for Needy Families Program*. In Robert A. Moffitt ed *Means-Tested Transfer Programs in the United States*, the University of Chicago Press, 2003.
- Morris, P., G. Duncan, and E. Clark-Kauffman, *Child Well-Being in an Era of Welfare Reform: The Sensitivity of Transition in Development to Policy Change*, *Developmental Psychology*, 2005. 41(6): p. 919-932.
- Murnane, R.J., J.B. Willet, and F. Levy, *The Growing Importance of Cognitive Skills in Wage Determination*. *Review of Economics and Statistics*, 1995. 77(2): p. 251-266.
- Murray, C., *Family Formation*, in *The New World of Welfare*, R.M. Blank and R.E. Haskins, Editors. 2001.
- National Center for Educational Statistics (NCES), *Mapping 2005 State Proficiency Standards onto the NAEP Scales*, NCES 2007-482, June 2007.
- Pepper, J., *The Intergenerational Transmission of Welfare Receipt: A Nonparametric Bounds Analysis*, *Review of Economics and Statistics*, 2000. 82(3): p. 472-488.
- Rivkin, S.G., E.A. Hanushek, and J.F. Kain, *Teachers, Schools, and Academic Achievement*. *Econometrica*, 2005. 73(2): p. 417-458.
- Ruhm, C.J., *Parental Employment and Child Cognitive Development*. *Journal of Human Resources*, 2004. 39(1): p. 155-92.
- United States Department of Health and Human Services, *Setting the Baseline: A Report on State Welfare Waivers*. 1997. Available at <http://aspe.hhs.gov/hsp/isp/waiver2/title.htm>.
- Zhang, L., *Political economy of income distribution dynamic*, *Journal of Development Economics*, forthcoming.
- Ziliak, J.P., D.N. Figlio, E.E. Davis, and L.S. Connolly, *Accounting for the Decline in AFDC Caseloads: Welfare Reform or the Economy?* *Journal of Human Resources*, 2000. 35(3): p. 570-86.

Table 1: National Welfare Reform and 4th Grade Math Scores

	1	2	3	4	5	6
SLE*Year = 2000	0.571 [0.736]	0.838 [0.725]	1.087 [0.739]	0.349 [0.817]	0.350 [0.819]	1.757 [0.752]*
SLE*Year = 2003	2.053 [0.425]**	2.334 [0.553]**	2.496 [0.558]**	2.012 [0.702]**	2.012 [0.700]**	2.54 [0.933]**
SLE*Year = 2005	2.44 [0.515]**	2.963 [0.886]**	3.142 [0.888]**	2.466 [0.854]**	2.474 [0.834]**	3.842 [1.240]**
Year = 2000	3.829 [0.721]**	2.896 [1.085]*				
Year = 2003	13.145 [0.828]**	12.118 [1.544]**				
Year = 2005	16.458 [0.845]**	13.25 [2.084]**				
White	15.066 [0.535]**	15.086 [0.532]**	15.013 [0.557]**	15.065 [0.549]**	15.065 [0.549]**	12.082 [0.432]**
Male	1.36 [0.250]**	1.367 [0.252]**	1.368 [0.266]**	1.372 [0.266]**	1.372 [0.266]**	2.042 [0.178]**
SLE	-17.838 [0.498]**	-17.766 [0.513]**	-17.876 [0.551]**	-16.502 [5.816]**	-16.452 [5.760]**	-35.756 [6.555]**
Accountability		2.283 [0.766]**				
Years of Accountability		0.182 [0.202]				
SLE*Accountability		-0.307 [0.786]	-0.503 [0.834]	0.014 [0.658]	0.003 [0.649]	-0.475 [0.507]
SLE*(Yrs of Accountability)		-0.055 [0.104]	-0.037 [0.110]	0.077 [0.120]	0.078 [0.122]	0.145 [0.126]
(Share Near Cutoff)*Year=2005					-0.005 [0.052]	
Constant	209.802 [0.673]**	153.685 [35.855]**	209.283 [0.439]**	191.078 [3.211]**	193.774 [3.183]**	190.277 [27.35]**
Observations	1018	1018	1018	1018	1018	515367

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

Note: The unit of observation in regressions 1-5 is a race*sex*low-income*state*year cell. The base year is 1996. Only states without major state-wide welfare reforms in place prior to 1995 are included in the sample. All columns include state fixed effects; Column 2 adds time-varying controls; Column 3 adds state*year fixed effects; Column 4 adds interactions between the *SLE* indicator and time-varying controls; and Column 5 controls for NCLB intensity, measured as share of students in each income group whose math test score in state tests in 2003 fell into either proficient or basic categories (Share Near Cutoff). Column 6 uses micro-data to isolate effects for FLE children. Some coefficient suppressed for readability.

Table 2: Heterogeneous Treatment Effects by Position on the Test Score Distribution, Matched by Mean 1996 Scores

<i>SLE Score Percentile</i>	<i>25th</i>	<i>Median</i>	<i>75th</i>	<i>90th</i>
<i>Mean Score in 1996</i>	189.64	208.46	226.31	241.86
<i>SLI Score Percentile</i>	<i>10th</i>	<i>25th</i>	<i>Median</i>	<i>75th</i>
<i>Mean Score in 1996</i>	192.42	209.67	228.16	245.83

Panel A: Basic Set of Controls				
SLE*Year = 2000	0.826 [0.920]	0.364 [0.849]	0.18 [0.691]	0.195 [0.658]
SLE*Year = 2003	1.83 [0.658]**	1.369 [0.631]*	1.265 [0.513]*	1.003 [0.550]+
SLE*Year = 2005	2.359 [0.858]**	2.137 [0.844]*	2.338 [0.760]**	2.394 [0.868]**

Panel B: Full Set of Controls				
SLE*Year = 2000	0.519 [0.955]	0.184 [0.869]	-0.054 [0.730]	-0.115 [0.705]
SLE*Year = 2003	1.676 [0.789]*	1.328 [0.727]+	1.132 [0.706]	0.655 [0.692]
SLE*Year = 2005	2.004 [0.841]*	1.981 [0.724]**	2.027 [0.710]**	1.786 [0.812]*

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

Note: The “basic” and “full” specifications correspond to the specifications in Column 2 and Column 4 of Table 1 respectively. See notes in Table 1.

Table 3: Heterogeneous Treatment Effects by Sex and Race**Panel A: Sex**

<i>Controls</i>	<i>Basic</i>		<i>Full</i>	
Sex	Girls	Boys	Girls	Boys
SLE*Year = 2000	0.567	1.036	0.162	0.584
	[0.844]	[0.894]	[0.991]	[1.028]
SLE*Year = 2003	2.007	2.751	1.884	2.359
	[0.791]*	[0.618]**	[1.123]+	[0.843]**
SLE*Year = 2005	2.744	3.401	2.602	2.759
	[1.075]*	[0.932]**	[1.233]*	[1.103]*

Panel B: Race

<i>Controls</i>	<i>Basic</i>		<i>Full</i>	
Race	White	Nonwhite	White	Nonwhite
SLE*Year = 2000	1.037	-0.722	0.016	1.307
	[0.631]	[1.539]	[0.781]	[1.438]
SLE*Year = 2003	2.418	1.476	1.356	3.786
	[0.526]**	[1.158]	[0.691]+	[1.286]**
SLE*Year = 2005	3.55	0.813	1.863	3.349
	[0.848]**	[1.545]	[1.004]+	[1.453]*

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

Note: The “basic” and “full” specifications correspond to the specifications in Column 2 and Column 4 of Table 1 respectively. See notes in Table 1.

Table 4: Welfare Reform Timing and 4th Grade Math Scores

	1	2	3	4	5
SLE*Months Since Reform	0.019 [0.005]**	0.019 [0.008]*	0.020 [0.008]*	0.015 [0.008]+	0.047 [0.015]**
Months Since Reform	0.043 [0.076]	0.041 [0.064]			
White	15.128 [0.480]**	15.118 [0.480]**	15.046 [0.509]**	15.181 [0.483]**	15.104 [0.514]**
Male	1.499 [0.220]**	1.504 [0.221]**	1.501 [0.234]**	1.508 [0.227]**	1.503 [0.241]**
SLE	-17.924 [0.504]**	-17.951 [0.500]**			
Accountability		1.622 [0.670]*			
Years of Accountability		0.298 [0.199]			
SLE*Accountability		0.119 [0.691]	-0.026 [0.730]	0.438 [0.574]	0.209 [0.564]
SLE*YrsAccount		-0.016 [0.103]	-0.008 [0.109]	0.001 [0.131]	0.027 [0.130]
Constant	209.835 [0.641]**	209.128 [36.162]**	209.251 [0.445]**	211.982 [37.677]**	198.962 [4.104]**
Observations	1293	1293	1293	1293	1293

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

Note: The unit of observation in each regression is a race*sex*free-lunch*state*year cell. All states are included in the sample. Other notes are the same as Table 1, except for Column 5, which includes state*free-lunch interactions.

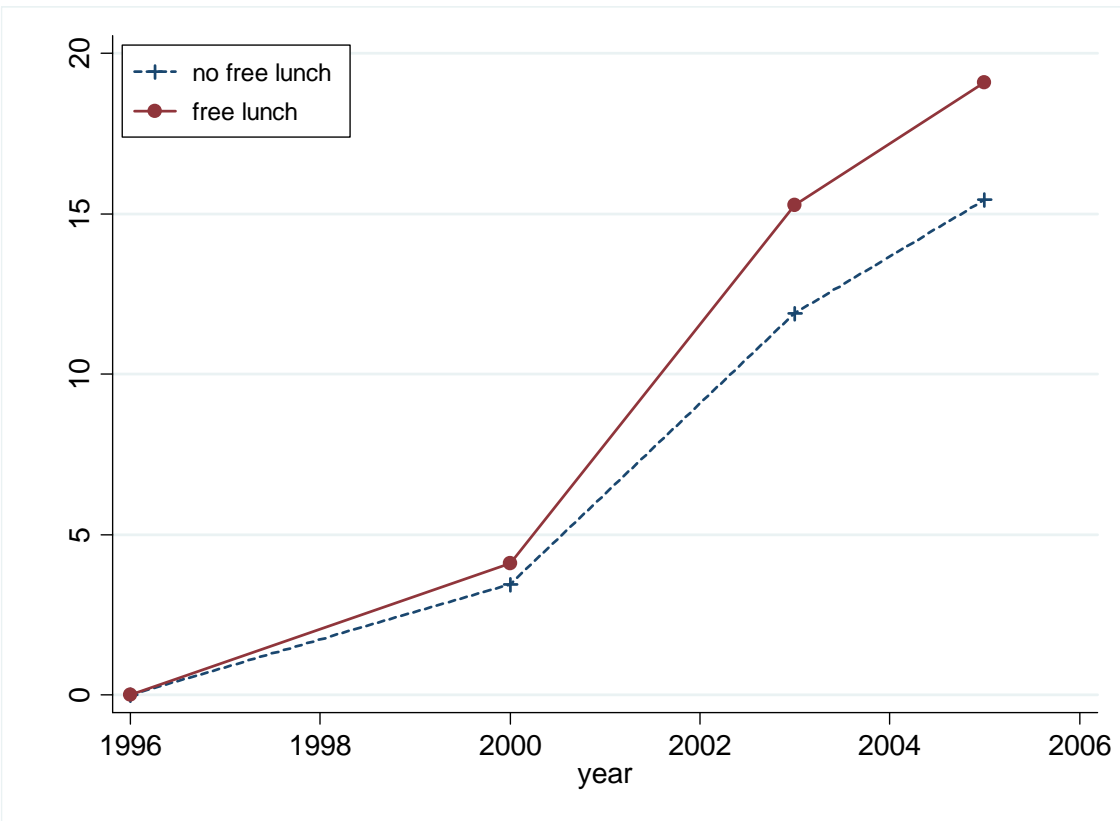
Table 5: Welfare Sanction Rules and 4th Grade Math Scores

	1	2	3	4	5
SLE*MS Any Sanctions	0.015 [0.009]+	0.016 [0.010]	0.017 [0.010]+	0.015 [0.011]	0.039 [0.016]*
SLE*MS G or F Sanctions	0.006 [0.012]	0.007 [0.013]	0.007 [0.013]	0.002 [0.012]	0.000 [0.010]
SLE*MS F Sanctions	0.007 [0.012]	0.007 [0.012]	0.006 [0.013]	-0.002 [0.009]	0.004 [0.014]
Months Any Sanctions	-0.106 [0.066]	-0.127 [0.055]*			
Months G or F Sanctions	0.000 [0.018]	-0.004 [0.018]			
Months F Sanctions	0.022 [0.017]	0.023 [0.016]			
White	15.128 [0.481]**	15.105 [0.480]**	15.046 [0.509]**	15.086 [0.503]**	15.105 [0.514]**
Male	1.501 [0.221]**	1.507 [0.223]**	1.502 [0.234]**	1.504 [0.236]**	1.504 [0.241]**
SLE	-17.937 [0.484]**	-17.958 [0.478]**	-17.983 [0.512]**		
Accountability		1.497 [0.619]*			
Years of Accountability		0.406 [0.199]*			
SLE*Accountability		0.022 [0.682]	-0.083 [0.719]	0.391 [0.632]	0.285 [0.582]
SLE*YrsAccount		-0.038 [0.119]	-0.034 [0.126]	0.158 [0.109]	0.03 [0.132]
SLE*HS Graduates				0.267 [0.087]**	0.034 [0.179]
SLE*Income				-7.714 [2.809]**	-1.609 [8.452]
SLE*Pupil-Teacher Ratio				-0.244 [0.170]	0.778 [0.607]
Constant	189.792 [34.853]**	200.642 [32.311]**	209.249 [0.438]**	178.886 [4.731]**	180.674 [3.427]**
Observations	1293	1293	1293	1293	1293

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

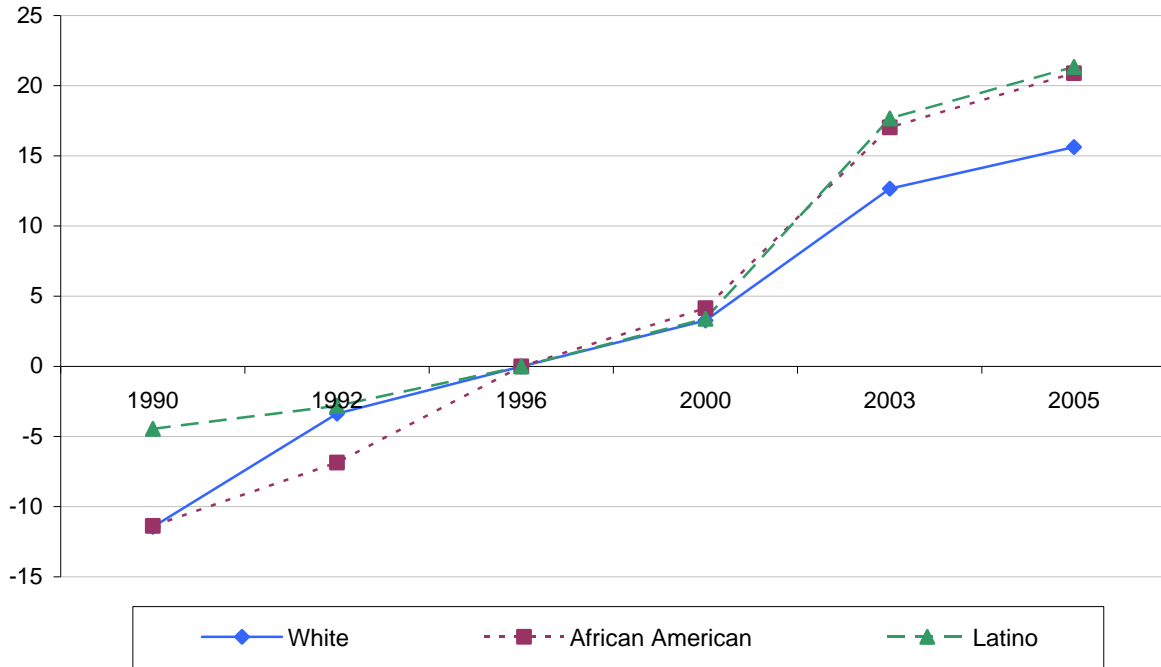
Note: Same as Table 4.

Figure 1: 4th Grade Math Score Gains Relative to 1996 Values by Eligibility for Free and Reduced-Price Lunches



Sources: Authors' calculation from NAEP. The values for 2000, 2003, and 2005 are test score gains relative to the 1996 value within each SLE category. Test scores are measured on a 500-point scale.

Figure 2: 4th Grade Math Score Deviations from 1996 Values by Race



Source: Authors' calculation from NAEP. The values for 1990, 1992, 2000, 2003, and 2005 are test score deviations from the 1996 value within each race. Test scores are measured on a 500-point scale.

Appendix: Robustness Check for Changes in Income Reporting

As discussed in Section 2, for the DD methodology to produce a consistent estimate of welfare reform's effect on low-income children's academic performance, it is necessary that the low-income and higher-income groups under comparison have stable compositions over time. Considering the modest income gains of welfare recipients and our inclusive criterion of 185% of federal poverty line, the low-income group (SLE) likely contains past, current, and potential welfare eligible children in all years, while the higher-income group (SLI) likely excludes them.¹⁸ Therefore, we argue that children leaving welfare as an outcome of reform are unlikely to switch status from SLE to SLI; welfare reform itself should not cause unstable group composition. Rather, our concern relates to changes in composition caused by changes in income reporting, which we address in this section.

The NAEP data provide test scores for three groups of students, based on the national school lunch program: eligible (SLE), ineligible (SLI), and information not available (NA). Students in the NA group were excluded from the analysis up to this point. They comprise a mix of eligible and ineligible students. If non-reporting is a random event, the exclusion of NA students will not bias our results. However, if non-reporting is systematically related to math performance and the rates of non-reporting are changing over time, our main findings of the previous sections may over- or under-estimate the true treatment effect.

In the NAEP data, the population shares for whites are stable at the start of the panel, between 1996 and 2000. Then, the shares in both SLE and SLI groups increase substantially between 2000 and 2003, and to a smaller extent between 2003 and 2005, while the percentage of students in the NA group drops dramatically from 14% to below 4% and then 2%. For non-whites, there is a large increase in the SLE group from 64% to 71%, a negligible increase in SLI from 26% to 27%, and a dramatic decrease in NA from 10% to 2%, over the decade.¹⁹

The steadily increasing trend in reported subsidized-lunch eligibility does not conform to the pattern of poverty rate movement, which declined from 11% to 8.7% between 1996 and 2000 and then increased to just over 10% in 2004,²⁰ suggesting that the apparent

¹⁸ Rough estimates of year 2000 income for single mothers who are former welfare recipients fall between 105% and 120% of the federal poverty level, including income from the EITC (Haskins 2001).

¹⁹ Year 2000 data for non-whites are not included in the table because reporting standards were not met.

²⁰ Source for poverty rate data is <http://www.census.gov/hhes/www/poverty/poverty.html>.

compositional shifts stem largely from improved reporting of eligibility. As a result of these changes, the income groups have an unstable composition in the sense that the same student may have been assigned to NA in an early year (and excluded) and assigned to SLE or SLI in a later year. School accountability and NCLB are the likely cause of improved reporting for both income groups. However, since schools must meet adequate yearly progress (AYP) within each category, there are no obvious incentives for principals to disproportionately increase SLE shares, as they will not generally lead to a relaxation of standards.

The potential size and direction of the bias in our main estimates of the treatment effect (Table 3) depends on the performance distributions of students in the three income groups: SLE, SLI, and NA. To illustrate, consider a simple difference-in-differences estimate with two time periods and three income reporting groups. Let P^0_{SLE} , P^0_{SLI} , and P^0_{NA} denote the observed test scores for each group in the initial (pre-welfare reform) period. The NA group is comprised of two sub-groups: SLE*, and SLI*. If schools had full information about students' family incomes, the true test scores of the SLE and SLI groups would be P^0_{SLE,SLE^*} and P^0_{SLI,SLI^*} . In the period following welfare reform, there is full information regarding income, and we observe P^1_{SLE,SLE^*} and P^1_{SLI,SLI^*} . According to our model (ignoring the independent stochastic error terms),

$$P^1_{SLE,SLE^*} = P^0_{SLE,SLE^*} + t + x, \text{ and}$$

$$P^1_{SLI,SLI^*} = P^0_{SLI,SLI^*} + t,$$

where t is the common time trend for both SLE and SLI groups, and x is treatment effect of welfare reform. Our estimate of the treatment effect is:

$$\begin{aligned} x_{obs} &= (P^1_{SLE,SLE^*} - P^1_{SLI,SLI^*}) - (P^0_{SLE} - P^0_{SLI}) \\ &= x + [(P^0_{SLE,SLE^*} - P^0_{SLE}) - (P^0_{SLI,SLI^*} - P^0_{SLI})] \end{aligned}$$

Therefore, the bias ($x_{obs} - x$) depends on the relative difference in baseline performance gaps between SLE and SLE*, and SLI and SLI*.

If assignment to NA in the initial period is random (uncorrelated with performance) within each of the two income groups, then $P^0_{SLE^*} = P^0_{SLE}$ and $P^0_{SLI^*} = P^0_{SLI}$, and x_{obs} is an unbiased estimate of x . The free lunch eligible and free lunch ineligible groups grow larger in the second period, but their relative performance is not altered. If, instead, the underlying ability difference between the SLE* group and the SLE group was more positive than the ability

difference between the SLI* group and the SLI group, then the bias term is positive, and our previous results are overestimates. For example, this would occur under the plausible situation that students with missing income information scored closer to the overall mean: if SLE* performed better than SLE, and SLI* worse than SLI. Although we never observe SLE* and SLI* separately, there is clear evidence in the data against random assignment: for whites in 1996, average test scores for the NA group are higher than scores for either SLE or SLI. Hence, we must consider the possibility that our estimates are biased upwards, and that the apparent gains from welfare reform are spurious and driven by compositional changes.

To test the robustness of our estimates, we first confirm that test score changes are not systematically related to changes in income reporting. We calculate the cross-state correlation between test score changes for low-income students (using both absolute and relative measures) between 1996 and 2005 and changes in the SLE share of students during the same time period. The pair-wise correlations are inconsistent in sign and never significant at conventional levels (lowest P-value of 0.44).

Next, we compute conservative lower-bounds on the treatment effects. We conduct a counter-factual experiment in which we assume that the NA group consists entirely of low-income students. We combine test scores from NA and SLE groups to produce the average pre-reform performance of the low-income group ($P^0_{SLE,SLE*,SLI*}$).²¹ The counter-factual estimate of the treatment effect is:

$$\begin{aligned} x_{cf} &= (P^1_{SLE,SLE*} - P^1_{SLI,SLI*}) - (P^0_{SLE,SLE*,SLI*} - P^0_{SLI}) \\ &= x + [(P^0_{SLE,SLE*} - P^0_{SLE,SLE*,SLI*}) - (P^0_{SLI,SLI*} - P^0_{SLI})] \end{aligned}$$

In our data, the NA group has higher average scores than the SLE group, which implies that $P^0_{SLE,SLE*,SLI*} > P^0_{SLE}$, and $x_{cf} < x_{obs}$.

The bias ($x_{cf} - x$) depends on the ability of SLI* students relative to the other groups, and the size of the SLI* group as a fraction of the NA group (f_{SLI}). Using the observed 1996 test scores and population shares, it is possible to determine the range of f_{SLI} for which the bias will always be negative under various assumptions about SLI* ability, and estimate an upper-bound on the magnitude of the positive bias. If SLI* baseline performance is at least as

²¹ In this example, the NA group only exists in the first period. In the data, the NA group is present throughout. For the counter-factual exercise, NA scores are included in the SLE average for all years, although, naturally, they constitute a shrinking share of the population of the group.

good as SLI performance ($P^0_{SLI*} \geq P^0_{SLI}$), the bias will be negative for all possible values of f_{SLI} . Under the far weaker assumption that students eligible for free or reduced-price lunches in the NA category score no better than those ineligible and also in the NA category ($P^0_{SLI*} \geq P^0_{SLE*}$), we can compute a maximum value for the bias as a function of f_{SLI} . For whites and the entire student population together, the bias is negative for the full range from zero to one. For non-whites, the bias is negative for values of f_{SLI} greater than 14%. For lower values of f_{SLI} , the bias ranges from zero to a maximum value of 0.0047. Hence, under the weak assumption that $P^0_{SLI*} \geq P^0_{SLE*}$, x_{cf} will be a lower-bound estimate of the treatment effect for whites and all students, and $x_{cf} - 0.0047$ will be a lower-bound estimate of the treatment effect for non-whites.

This measure ignores information about the actual shares of non-white students ineligible for reduced price lunches; among non-white students with a recorded eligibility status, the share ineligible ranges across years from 27.6% to 28.9%, well above the 14% cutoff to ensure negative values for $(x_{cf} - x)$. In addition, note that these are values for the extreme case in which $P^0_{SLI*} = P^0_{SLE*}$. In the more likely scenario that $P^0_{SLI*} > P^0_{SLE*}$, the bias will be more negative.

Appendix Table 5 shows the estimates produced by assigning the entire NA group to the SLE category and repeating the estimation of the previous sections with the new data. What we should focus on here is the fact that the coefficients for the effects of welfare reform remain positive. In the basic specification, including controls for accountability reform, found in Column 2, $\beta_{SLE,2000} = 0.502$, $\beta_{SLE,2003} = 1.35$ (significant at 5%), and $\beta_{SLE,2005} = 1.25$. Hence, even our *lower-bound* point estimates are positive, although not always statistically significant. The counterfactual estimates for β_{SLEMSR} are smaller than the original estimates (found in Table 4), and range from 0.006 to 0.015. While they are not statistically significant, they are most likely under-estimates, and the positive point estimates support the positive finding.

Appendix Table 1: 4th Grade Math Score Summary Statistics

Panel A: By School-lunch Eligibility

Year		<i>All</i>	<i>SLI</i>	<i>SLE</i>
1996	Mean	217.08	227.56	207.83
	St. Dev.	14.42	9.28	11.52
2000	Mean	220.68	231.01	211.93
	St. Dev.	13.77	8.98	10.73
2003	Mean	231.04	239.46	223.10
	St. Dev.	12.04	8.40	9.23
2005	Mean	234.70	243.00	226.92
	St. Dev.	11.71	7.85	9.11

Panel B: By School-lunch Eligibility and Position on Test Score Distribution

year		<i>SLI</i>					<i>SLE</i>				
		10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
1996	Mean	192.42	209.67	228.16	245.83	261.36	172.24	189.64	208.46	226.31	241.86
	St. Dev.	11.42	10.84	9.77	8.71	8.09	12.37	12.07	11.83	11.41	10.73
2000	Mean	196.83	214.04	231.94	248.89	263.56	177.21	194.68	212.94	230.23	245.25
	St. Dev.	11.02	10.14	9.05	7.97	7.42	12.70	11.69	10.90	10.16	9.67
2003	Mean	206.49	222.72	240.22	256.98	271.16	190.69	206.62	223.80	240.40	254.61
	St. Dev.	10.59	9.85	8.69	7.69	7.14	10.07	9.93	9.51	9.08	8.57
2005	Mean	209.86	226.36	243.65	260.21	275.03	193.85	210.43	227.89	244.43	258.58
	St. Dev.	10.39	9.23	8.05	7.08	6.50	10.19	9.93	9.42	8.67	8.11

Panel C: By School-lunch Eligibility, Race, and Gender

Year		<i>SLI</i>				<i>SLE</i>			
		White Boy	White Girl	Non-White Boy	Non-White Girl	White Boy	White Girl	Non-White Boy	Non-White Girl
1996	Mean	234.47	232.14	217.50	217.88	217.71	216.29	195.97	197.69
	St. Dev.	4.74	4.12	7.05	6.24	5.57	4.82	5.97	6.08
2000	Mean	237.30	234.80	221.24	221.57	220.60	219.33	201.74	202.03
	St. Dev.	4.15	4.03	8.14	7.43	5.43	4.62	7.52	5.98
2003	Mean	246.94	243.86	232.93	231.63	232.10	229.36	215.45	214.51
	St. Dev.	4.00	3.89	6.64	6.00	4.18	3.39	5.98	5.02
2005	Mean	249.82	246.74	236.71	236.38	235.44	233.06	219.64	218.62
	St. Dev.	4.17	4.12	6.27	6.05	4.28	4.35	5.86	5.46

Source: Authors' calculation from National Assessment of Educational Progress (NAEP).

Appendix Table 2: Welfare Reform and School Accountability Data

State	Earliest State-Wide Waiver Implemented	TANF Implemented	Initial Sanction Date	Initial Sanction Type	Initial or Sanction Date	G or F Sanction Type	Starting date of FSanction type	End of FSanction Type	Consequential Ability Introduced	Report-Card Ability Introduced
Alabama		Nov-96	Nov-96	P	May-97	G		G	1997	
Alaska		Jul-97	Jul-97	P	Jul-02	G		G		2001
Arizona	Nov-95	Oct-96	Nov-95	P	Aug-97	G		G		2000
Arkansas	Jul-94	Jul-97	Jul-97	P	Aug-01	G		G	1999	
California	Dec-92	Jan-98	Jan-98	P				P	1999	
Colorado		Jul-97	Jul-97	G	Jul-97	G		G		2002
Connecticut	Jan-96	Oct-96	Jan-96	G	Jan-96	G		G	1993	
Delaware	Oct-95	Mar-97	Mar-97	P				P	1998	
DC		Mar-97	Mar-97	P	Jul-97	G		G		1997
Florida		Oct-96	Oct-96	F	Oct-96	F	Oct-96	F	1999	
Georgia	Jan-94	Jan-97	Jan-94	P	Jul-97	G		G	2000	
Hawaii	Feb-97	Jul-97	Jul-97	P	Aug-99	F	Aug-99	F		2001
Idaho		Jul-97	Jul-97	F	Jul-97	F	Jul-97	F		1997
Illinois	Nov-93	Jul-97	Oct-95	P	Jul-97	G		G		1999
Indiana	May-95	Oct-96	May-95	P	May-03	G		G		1995
Iowa	Oct-93	Jan-97	Oct-93	F	Oct-93	F	Oct-93	F	2003	
Kansas		Oct-96	Oct-96	F	Oct-96	F	Oct-96	F		1995
Kentucky		Oct-96	Oct-96	P	Nov-98	G		G	1995	
Louisiana		Jan-97	Jan-97	P	Mar-98	G		G	1999	
Maine		Nov-96	Nov-96	P				P		1999
Maryland	Mar-96	Dec-96	Oct-96	F	Oct-96	F	Oct-96	F	1999	
Massachusetts	Nov-95	Sept-96	Nov-95	G	Nov-95	G		G	1998	
Michigan	Oct-92	Sept-96	Oct-94	G	Oct-94	G	Apr-97	F	1998	

State	Earliest Wide Imple- mented	State- Waiver TANF Implemented	Initial Sanction Date	Initial Sanction Type	Initial or Sanction Date	G F Sanction Type	Starting date of type	End FSanction Type	Consequential Account- ability duced	Report-Intro-ity Card-ity Accountabil- Introduced
Minnesota		Jul-97	Jul-97	P	Jul-01	G		G		1996
Mississippi	Oct-95	Jul-97	Jul-97	F	Jul-97	F	Jul-97	F		1994
Missouri	Jun-95	Dec-96	Jun-95	P				P		1997
Montana	Feb-96	Feb-97	Feb-97	P	May-98	G		G		1998
Nebraska	Oct-95	Dec-96	Dec-96	F	Dec-96	F	Dec-96	F		2001
Nevada		Dec-96	Dec-96	P	Jan-98	G		G	1996	
New Hampshire		Oct-96	Oct-96	P				P		1993
New Jersey	Oct-92	Jul-97	Oct-92	P	Jul-97	G		G	1997	
New Mexico		Jul-97	Jul-97	G	Jul-97	G		G	2003	
New York		Nov-97	Nov-97	P				P	1998	
North Carolina	Jul-96	Jan-97	Jul-96	P	Dec-99	G		G	1993	
North Dakota		Jul-97	Jul-97	P	Jul-98	G		G	2003	
Ohio	Jul-96	Oct-96	Jul-96	G	Jul-96	G		G	1997	
Oklahoma		Oct-96	Oct-96	F	Oct-96	F	Oct-96	F	1996	
Oregon	Feb-93	Oct-96	Jul-96	G	Jul-96	G		G	2000	
Pennsylvania		Mar-97	Mar-97	G	Mar-97	G		G		1999
Rhode Island		May-97	May-97	P				P	1997	
South Carolina		Oct-96	Oct-96	F	Oct-96	F	Oct-96	F	1999	
South Dakota	Jun-94	Dec-96	Jun-94	P	Sept-97	G		G		1997
Tennessee	Sept-96	Oct-96	Sept-96	F	Sept-96	F	Sept-96	F	1996	
Texas	Jun-96	Nov-96	Jun-96	P				P	1994	
Utah	Jan-93	Oct-96	Jul-96	G	Jul-96	G		G	2003	
Vermont	Jul-94	Sept-96	Jul-94	P				P	1999	
Virginia	Jul-95	Feb-97	Feb-97	F	Feb-97	F	Feb-97	F	1998	
Washington	Jan-96	Jan-97	Jan-97	P						1998

State	Earliest Wide Imple- mented	State- Waiver TANF Implemented	Initial Sanction Date	Initial Sanction Type	Initial or Sanction Date	G F Sanction Type	Starting date type	End of FSanction Type	Consequential Account- duced	Report-Card Intro-ity Introduced
West Virginia	Feb-96	Jan-97	Feb-96	P	Jan-97	G		G	1997	
Wisconsin	Jan-96	Sept-97	Jan-96	G	Jan-96	G		G	1993	
Wyoming		Jan-97	Jan-97	G	Aug-97	F	Aug-97	F		1999

Source: Crouse (1999) for the dates of state-wide waiver implementation and TANF implementation. U.S. DHHS (1997), and the Urban Institute Welfare Rules Database for date and severity of state sanction policies. Hanushek and Raymond (2005), Fletcher and Raymond (2002), Goertz and Duffy (2001), and various state Department of Education websites for the introduction dates of state accountability systems.

Note: Sanction types: P = partial loss of benefits, F = full loss of benefits, G = initial penalty is a partial loss of benefits, maximum penalty is a total loss.

Appendix Table 3: Control Variable Summary Statistics

<i>Year</i>		<i>Pupil-Teacher Ratio (K-12)</i>	<i>Per Pupil Expenditure, (K-12, \$)</i>	<i>Population % HS Diploma (age ≥ 25)</i>	<i>Population % College Degree (age ≥ 25)</i>	<i>Average Income (\$)</i>	<i>Unemployment Rate</i>
1996	Mean	16.8	3,663.8	79.1	21.1	14,156.6	5.9
	Min	13.3	2,293.2	67.8	11.7	10,232.6	3.1
	Max	24.3	6,094.3	88.4	35.6	19,961.3	9.7
2000	Mean	16.2	3,860.9	82.1	22.8	14,818.5	5.5
	Min	13.1	2,498.5	72.7	13.7	11,062.0	2.8
	Max	22.9	6,099.4	90.4	36.6	21,011.2	8.6
2003	Mean	15.6	4,231.0	83.9	24.2	15,621.3	4.8
	Min	12.0	2,676.3	75.8	14.9	11,722.1	2.9
	Max	21.9	6,887.5	91.5	38.3	22,420.0	7.5
2005	Mean	15.5	4,436.8	84.8	25.2	16,140.9	4.8
	Min	11.7	2,728.3	77.6	15.6	12,098.6	3.1
	Max	22.0	7,181.3	91.3	40.3	23,803.1	7.3

Source: Authors' calculation from U.S. Census and the U.S. Department of Education.

Note: Controls are for regressions for 4th grade math scores. All dollar values are in constant 1983 dollars. Pupil-teacher ratio and expenditure per student are averages over the 1992-1995, 1996-1999, 1999-2002, and 2001-2004 periods for test years 1996, 2000, 2003, and 2005 respectively; other variables are averages over the 1987-1995, 1991-1999, 1994-2002, and 1996-2004 periods.

Appendix Table 4: Welfare Reform and 8th Grade Math Scores

	1	2	3	4
Controls	Basic	Full	Basic	Full
SLE*Year = 2000	-0.461 [0.890]	-0.207 [0.936]		
SLE*Year = 2003	0.233 [0.824]	0.849 [0.921]		
SLE*Year = 2005	2.087 [1.046]+	2.359 [1.307]+		
SLE*Months Since Reform			0.003 [0.009]	0.002 [0.010]
Months Since Reform			0.046 [0.088]	
Year = 2000	0.396 [1.449]		-2.142 [4.079]	
Year = 2003	3.827 [1.891]+		-0.382 [7.379]	
Year = 2005	2.495 [2.445]		-1.608 [9.479]	
White	19.754 [0.759]**	19.64 [0.793]**	19.93 [0.711]**	19.794 [0.749]**
Male	0.686 [0.226]**	0.677 [0.240]**	0.668 [0.224]**	0.636 [0.232]**
SLE	-17.202 [0.607]**		-17.673 [0.544]**	
Accountability	2.491 [0.806]**		1.156 [0.784]	
Years of Accountability	0.606 [0.278]*		0.476 [0.271]+	
SLE*Accountability	-2.221 [0.963]*	-1.558 [0.806]+	-1.086 [0.806]	-0.633 [0.677]
SLE*YrsAccount	-0.096 [0.125]	0.014 [0.129]	-0.08 [0.120]	0.133 [0.132]
Constant	261.96 [58.664]**	214.818 [4.116]**	221.246 [42.798]**	228.397 [4.140]**
Observations	982	982	1239	1239
R-squared	0.93	0.95	0.93	0.95

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

Note: The first two columns estimate Equation 2 and have the same specifications as those in Columns 2 and 4 of Table 3; the last two columns estimate Equation 3 and have the same specifications as those in Columns 2 and 4 of Table 6. See notes in Table 3.

Appendix Table 5: Lower-Bound Estimates of Treatment Effects for 4th Grade Math Scores

	1	2	3	4
SLE*Year = 2000	0.233 [0.799]	0.502 [0.822]	0.708 [0.833]	-0.102 [0.885]
SLE*Year = 2003	1.077 [0.547]	1.35 [0.622]*	1.534 [0.616]*	0.881 [0.764]
SLE*Year = 2005	0.817 [0.660]	1.25 [0.906]	1.399 [0.893]	0.437 [0.930]
Observations	1038	1038	1038	1038
R-squared	0.93	0.93	0.95	0.95

	5	6	7	8
SLE*Months Since Reform	0.006 [0.006]	0.008 [0.007]	0.009 [0.008]	0.004 [0.007]
Observations	1319	1319	1319	1319
R-squared	0.93	0.93	0.95	0.95

Robust standard errors clustered at state level in brackets. +: significant at 10%; *: significant at 5%; **: significant at 1%.

Note: The unit of observation in each regression is a race*sex*free-lunch*state*year cell. The low-income group is a combined group of students recorded as eligible for free or reduced-price lunch and students with no recorded information on eligibility. The base year is 1996. For the top panel, only states without major state-wide welfare reforms in place prior to 1995 are included in the sample; other notes are the same as Table 3. All states are included for the bottom panel; other notes are the same as Table 6.