Did Welfare Reform Improve the Academic Performance of Children in Low-Income Households?

Amalia R. Miller[*] and Lei Zhang[†‡]

April 2007

Abstract:

During the 1990s, U.S. welfare policy underwent dramatic reforms aimed at promoting employment and reducing dependence on public assistance. While the immediate effects of welfare reform on adult employment and family income have been studied extensively, little is known about its long-run effect on the wellbeing of children in low-income families. This paper evaluates the impact of welfare reform on the academic performance of children in low-income families using nationally representative data on math achievement that spans a ten year period starting just prior to welfare reform. In a generalized difference-in-differences framework, we use children from more affluent families as a control group to simulate the unobservable counterfactual time trend. The key finding is an improvement in the relative math scores of fourth-grade students from low-income families, which is statistically significant for 2003 and 2005. No consistent differences are found by race, sex or ability. The treatment effect is smaller for eighth-grade students, but the benefits from early exposure appear to persist and even increase with age. Preliminary analysis suggests that these beneficial effects are associated with relative improvements in low-income children's time use and interaction with parents.

[*] Department of Economics, University of Virginia, Charlottesville, VA. E-mail: armiller@virginia.edu.
[†] Department of Economics, Clemson University, Clemson, SC. E-mail: zlei@clemson.edu.
[‡] We thank Doug Bernheim, William Dougan, Eric Hanushek, John Pepper, Macke Raymond, Sarah Turner and seminar participants at Clemson, Maryland, Virginia and the BLS for helpful comments, as well as Liliana Danila, Felicia Emery and Maria Fitzpatrick for excellent research assistance. Zhang thanks Clemson University College of Business and Behavioral Science for financial support in the form of Summer Research Grant.

1. Introduction

During the 1990s, U.S. states implemented a variety of reforms to their cash assistance programs to low-income families. These experiments culminated in fundamental national welfare reform legislation in August 1996, in the form of the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA), which replaced the Aid to Families with Dependent Children (AFDC) program with Temporary Assistance for Needy Families (TANF). The primary goal of these reforms was to promote employment among recipients and to reduce dependence on public assistance. To that end, the policy changes included job subsidies, work requirements, increased funding for child support, and time limits for receiving assistance. The importance of cash assistance programs for redistribution, the dramatic changes imposed by welfare reform, and the ongoing political debates about welfare policies, have together attracted scholars and policymakers to assess the impact of welfare reform on the well-being of individuals in poor households.

This paper addresses an important but largely unexplored long-term impact of welfare reform: how did welfare reform affect the academic performance of children in low-income families? Children are a vulnerable population that has attracted policymakers' particular interest.[1] Successful welfare reform will not only reduce poverty and improve the economic well-being of families in poverty, but will also help to break intergenerational links in dependency (Pepper 2000) by improving opportunities for children. Numerous theoretical and empirical studies have demonstrated that individuals with better education have higher future earnings (for example, Becker 1993, Zhang 2005) and health (Lleras-Muney 2004); therefore, one important criterion for a successful welfare policy is its beneficial impact on the educational outcomes of low-income children.[2]

To our knowledge, this study provides the first attempt to measure the effects of national and state-wide welfare reforms on the academic performance of low-income children. We em-

[1] Indeed, shielding children from the adverse consequences of household poverty is a primary goal of the welfare system in the United States. Childless adults with very low incomes are ineligible for AFDC/TANF or Medicaid (unless they qualify via disability or old age) and can claim only substantially reduced credits under the Earned Income Tax Credit (EITC).

[2] A controversial issue regarding public transfers is the relative importance of long-term versus short-term effects. An argument for education and skills training programs in the Family Support Act of 1988 was the potential to produce long-term increases in earnings for participants, which would offset the short-term reductions in income (Moffitt 2003).

ploy 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 change in academic performance for students in low-income families *caused* by welfare reform, we use more affluent students as a control group who were not directly affected by the policy changes to simulate the unobservable counterfactual time trend –what would have occurred to low-income students' test scores in the absence of welfare reform. The basic difference-in-differences (DD) framework is augmented by the inclusion of control variables for the economic and educational environment as well as additional fixed effects and interaction terms. In particular, we deal with confounding effects due to contemporaneous policy changes (expansion of EITC and adoption of school accountability programs) that may affect low-income and more affluent students differentially. The main approach evaluates the impact of national reform using states that had not adopted major reforms prior to 1995. A secondary approach uses all states, and exploits timing variation in the implementation of state-wide waivers and TANF programs.

We find no evidence that welfare reform harmed the academic performance of low-income students. Rather, the key finding is an improvement in the relative math scores of 4^{th}-grade students in low- income families, which is statistically significant for 2003 and 2005. Preliminary analysis suggests that these beneficial effects are associated with relative improvements in low-income children's time use and interaction with parents. This result suggests that the reduction in maternal time available for investment 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. We investigate heterogeneous treatment effects by student ability, sex, and race, and find little variation. While the effects for 4^{th} grade students are larger and more significant than those for 8^{th} grade students for the same duration of exposure to welfare reform, the 4^{th} grade gains appear to last through 8^{th} grade. In a value-added framework, welfare reform is also shown to improve test score growth between 4^{th} and 8^{th} grades.

Our study adds an important dimension to the welfare reform evaluation literature. The great bulk of the existing research has focused on outcomes for adults and households. Blank (2002), Moffitt (2003), and Grogger and Karoly (2005) provide excellent reviews of this extensive literature. 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, Figlio and Ziliak 1999, Hoynes 2000, Ziliak et al. 2000), although the TANF gains may be the result of a booming economy (Schoeni and Blank 2000). There is some controversy regarding the effects of welfare reform on family income. Poverty rates declined during the late 1990s, but the economic boom may have been the cause. 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" during the period (Haskins 2001). Meyer and Sullivan (2004) use a difference-in-differences strategy, comparing women according to their marriage and motherhood status, and show that consumption did not decline for vulnerable women as a result of welfare and tax reform. Other relevant outcomes relate to family structure, but evidence is not conclusive and the effects have not been demonstrated with national data (Horn and Sawhill 2001, Murray 2001).

The long-run effect of welfare reform has largely been ignored by scholars, in part due to the lack of high quality national panel data on children's outcomes spanning the relevant period. The most compelling studies to date employed an experimental framework with random assignment into treatment (welfare reform) and control (traditional welfare) groups, using data from five welfare demonstrations.[3] They find evidence of improved academic achievement for younger children in the treatment group, measured by self-reports, teacher-reports and test scores (Duncan and Chase-Lansdale 2001, Grogger and Karoly 2005).[4] Although proper randomization eliminates self-selection bias, experiments suffer from some well-documented limitations, which make it uncertain whether the findings would generalize to national data and actual reforms. The experiments were limited geographically and in the nature of their reforms; demonstrations tended to employ weaker work requirements than those in state-wide waivers and TANF programs.[5] By contrast, this paper measures the effects of state-wide and national welfare reforms

[3] 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.

[4] A recent observational study (Chase-Lansdale et al. 2003) using post-reform data on low-income populations finds little evidence that transitions out of welfare or into work are associated with outcomes for children or adolescents.

[5] Some additional limitations are listed in Grogger and Karoly (2005), for example: random assignment studies can be costly and difficult to implement properly, they miss initial entry effects on people at risk of welfare participation, their findings may not generalize to universal large-scale programs, their randomization may be tainted by "cross-over" and attrition, and participants may not correctly perceive which of the new policies apply to them.

on a nationally representative group of children from low-income families, including both current and previous welfare participants.

2. Conceptual Framework

School performance is generally modeled as the outcome of student ability, school inputs, and family inputs cumulative to the time of performance measurement (Haveman and Wolfe 1995, Hanushek 2002):

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

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 ν_{it} is a stochastic error term.

Welfare reform affects the school performance of low-income students through family inputs and student effort ($F_i^{(t)}$), combining direct and indirect effects. The direct channel involves students altering their behavior in response to the changed incentives for schooling due to the reduced availability and attractiveness of welfare as a long-term substitute for paid employment. This channel is likely negligible for males and for younger children, but it may be important for adolescent girls.

The indirect channel is changes in parental behavior and family environment caused by welfare reform. Among the important changes discussed above, maternal employment increased, welfare participation declined, and family income may increased, but not for all families. Maternal employment can hinder child cognitive development by reducing the available hours that can be devoted to home production, such as supervising and disciplining children, reading to them, and assisting with homework. On the other hand, working mothers may feel more secure and confident as a result of their work outside of the home, resulting in greater productivity at home. Working mothers may also provide their children with improved and more stable daily routines, 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.

Prior studies of the maternal employment and child human capital produce mixed evidence. Ruhm (2004) finds that maternal labor supply in early childhood is associated with diminished verbal, math and reading ability for pre-school aged children, after controlling for a variety of observable factors. These negative effects are limited to children in two-parent households, and may not be applicable to our population of interest. Blau and Grossberg (1992) find negative effects of maternal labor supply in the first year of life, but a zero cumulative effect for the first three years. Here, the harmful effects are concentrated among high-income households. Baum (2003) finds negative effects from time devoted to employment in the first year, which are offset by the increased income. Gregg et al. (2005), using a UK sample, find mixed effects of maternal employment on early childhood cognitive development, but the adverse effects are restricted to more affluent or better-educated mothers. These studies employ a variety of controls for family background, including family fixed effects (see sibling comparisons in Ruhm 2004). However, the endogeneity of the work decision, which is likely correlated with unmeasured household characteristics, is a major concern. Turning to income, Blau (1999) finds only trivial direct effects of family income on child development.

The different channels may operate in different directions, and the net effect of welfare reform on school performance is inherently an empirical question. Our approach in this paper is to uncover this net effect in a reduced-form analysis. In so doing, we do not intend to show the exact mechanisms through which welfare reform affects children's school performance. We therefore rewrite Equation (1) as:

$$P_{ijct} = f(F1_{ijc}^{(c,t)}, 1(j=1)\cdot F2_{ijc}^{(WR,t)}, S_{ijc}^{(t-g,t)}, A_i) + \nu_{ijct}, \qquad (1')$$

where j, c, g denote student i's income group, birth year cohort, and grade level respectively. j is equal to 1 for the low-income group and 0 for the higher-income group. The superscripts (c,t), (WR,t), and $(t-g,t)$ denote that the respective input is accumulated between year c and year t, between year of welfare reform implementation (WR) and year t, and between the year starting school ($t-g$) and year t.

Equation (1') incorporates two observations. First, family inputs take effect when a child is born, and school inputs directly affect school performance while a child is in school.[6] Second, as a thought experiment, we can decompose family inputs into two components: one part ($F1$) is not affected by welfare reform, the other part ($F2$) is affected by welfare reform. Welfare reform only affects the school performance of the low-income group, and it works through its effect on the family inputs ($F2$) from the time of welfare reform implementation to the time of outcome measurement. This specification thus hypothesizes that the length of time low-income families are exposed to welfare reform legislation matters: we can imagine that families with longer exposure to welfare reform will alter their behavior more in response to the reform (larger change in $F2$), and hence children in these families will be more affected. Note that there is likely a lag between welfare reform and family's response to the reform (lack of change in $F2$ in the immediate years following reform); therefore, children's performance may not be affected by welfare reform if the time of measurement is sufficiently close to the time of welfare reform implementation.

In the empirical analysis, we use the student group eligible for free or reduced-price lunch (FLE) as the low-income group and the student group ineligible for the school lunch program (FLI) as the higher-income group. The FLE group is a superset of the welfare eligible category, since income cutoffs are higher for the free and reduced price lunch program (185% of the federal poverty level).[7] Thus the FLE group is likely to contain past, current, and potential welfare eligible children, as well as children from somewhat wealthier families, and this group will be affected by welfare reform. At the same time, the FLI group is unlikely to contain welfare recipient children, and thus should not be directly influenced by welfare reform.

We adopt and estimate a linear specification of the relation in (1'), i.e.,

$$P_{ist} = \beta_s + \beta_{FLE} \cdot FLE_i + \beta_t \cdot \tau_t + \beta_{FLE,t} \cdot FLE_i \cdot \tau_t + \beta_X \cdot X_{ist} + \nu_{ist}, \qquad (2)$$

[6] It is possible that family income and other behavior before or during pregnancy may affect future school performance; this could work through their impact on a child's innate ability.

[7] Use federal poverty line as the cutoff income for welfare eligibility. In 2000, of all the families with children between 5 and 17 years of age and with income less than 185% of the federal poverty line, 45% have income below poverty line, and 26% have income below poverty line and are female-headed (Census, 2000).

where *s* indexes state, *t* indexes time, *i* indexes student race-sex-lunch eligibility group, $FLE_i = 1$ if i is the FLE group, and $\tau_t = 1$ if t is a post-reform year. In this difference-in-differences (DD) approach, each of the year-eligibility interaction terms ($\beta_{FLE,t}$) represents an estimate of the treatment effect on the FLE group from the year of welfare reform to that year, and corresponds to the net effect of $F2_{ijc}^{(WR,t)}$ in equation (1'). Other terms capture state fixed effects (β_s), year fixed effects (β_t), fixed differences between FLE and FLI students (β_{FLE}), and time-varying controls (β_X). X_{ist} includes proxies for a state's school inputs (S), measured as averages for the period a group is in school, and proxies for a state's overall trend in family inputs ($F1$), measured as averages between the birth year and the test year. To capture the idea that the FLE and FLI groups may face differential input conditions, or they may respond differentially to variations in aggregate economic or education conditions, we may also include interactive terms between FLE_i and the control variables in X_{ist}.

In all estimation that follows, test scores of higher-income students are used to create counterfactual trends for what would have happened to test scores of low-income students in the absence of welfare reform. $\beta_{FLE,t}$ is identified by within group variation over time, and the coefficient captures the post-reform deviation of the low-income group's performance from the high-income group's performance, after controlling for the mean changes in the performance observed over the period. If the FLE proxy for poverty is overly inclusive or there exist positive cross-income peer effects, the $\beta_{FLE,t}$ estimate will be attenuated.

One potential concern of this DD study is that even in the absence of policy changes under study, underlying trends in the outcome variable of interest could lead to spurious estimates of policy effects. This could arise for two reasons. First, other factors that differentially affect the performance of low-income and higher-income groups may be changing contemporaneously. Given the cumulative nature of the education production process, these confounding forces may be events that occurred before or after PRWORA. We identify two important policy changes that could potentially bias the effect of welfare reform on low-income students' school performance: (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. Second, the composition of the low-income and high-income groups is not stable over the period studied. This could arise because of either an exogenous change in the definition of income status or an endogenous change in family income as a result of the welfare reform. We address both sources of concern in the next sections.

3. Data

The unit of observation is a state-year-race-gender-income cell, where income is defined by eligibility for federal (USDA) free or reduced-price school lunches. Data used in the empirical analysis come from various sources.

3.1 School Performance Measures

School performance is primarily measured by the average mathematics test scores for a representative sample of about 3,000 fourth graders in each state.[8] These tests are conducted by the Department of Education through the National Assessment of Educational Progress (NAEP) program and are given between late January and early March of the test year. These tests provide an independent measure of knowledge and aptitude of the U.S. students in mathematics and have been widely used for cross-state comparison of student performance. The math test is on a 500-point scale. We use the mathematics test scores for all years in which NAEP reports state average test scores separately for the FLE and FLI groups: 1996, 2000, 2003, and 2005.[9] Therefore, the birth cohorts of the 4th graders are 1987, 1991, 1994, and 1996.

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

[9] The NAEP also tested fourth graders in reading and science. Unfortunately, the scores are only available by free lunch eligibility starting in 1998 and 2000, respectively, making it impossible to conduct a full empirical analysis on those subject areas. Using the limited time periods, and the same sets of fixed effects and control variables, we find some evidence that the gains in math were repeated in other subjects. Relative to FLI students, FLE students achieved greater growth in reading scores from 1998 to 2002, 2003 and 2005; the point estimates are universally positive, although not always significant statistically. Comparing science scores in the two available years (2000 and 2005), we also find relative gains for the FLE group, statistically significant in all but one specification. Scores for twelfth-grade tests are not available at the state level.

Panel A of Table 1 reports the mean and standard deviation of the state test scores for 4th graders in each test year. Over the ten-year period between 1996 and 2005, there is an across-the-board improvement in students' math test scores, manifested by the increase in the average test score and the decrease in the standard deviation of the test scores across states. The increase in mean test scores is more pronounced when we compare test scores in 1996 and 2000 with those in 2003 and 2005. From 1996 to 2005, 4th graders experience an increase in mean test score of more than 1 standard deviation; we observe similarly an increase in test scores and a decrease in the variation in the test scores for the FLE and FLI groups separately.[10]

The pattern of test score gains, however, differs substantially between the FLE and FLI groups; this difference is illustrated in Figure 1, which plots the average test score gains for each group relative to its 1996 average. The gains for the FLE group are larger than those for the FLI 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 FLE group experiences an even larger increase, which, given the timing of the welfare reform, may be attributed to the impact of welfare reform on family inputs to the education production process. The increasing gains by the FLE group over time also suggest the existence of a lag between welfare reform implementation and its impact on low-income children's test scores. Foreshadowing the exploration of heterogeneous effects below, Figures 2 and 3 further depict the test score gains for each ability and gender-race group.[11] For 4th graders, the difference between the FLE and FLI groups exists universally for all ability and gender-race groups, and appears slightly more pronounced for low-performers and non-whites. These figures are illustrative of the overall time trends, and correspond to regressions with time trends, FLE, and their interaction, but with no additional covariates. Estimates that follow include controls for changes in observable factors during the period, as well as more precise measures of the time elapsed since actual welfare reform. Panels B and C of Table 1 presents the mean and standard deviation of the test score for each ability and gender-race group by free-lunch eligibility.

[10] The overall math improvement trend between 1996 and 2004 is also found in the national "long-term trend" sample of the NAEP, with boys, girls, whites and non-whites each experiencing over 10 point gains in average scores.

[11] The NAEP does not report "unreliable" scores for cells in which the underlying sample size of test-takers is too small. In order to preserve geographic representation and a reasonable number of observations, we consider two races: white and non-white. Non-white includes all races other than white. The final 4th grade sample contains 1,309 observations, or 80% of the potential maximum.

3.2 Welfare Reform Variables

The first two columns of Table 2 show the earliest date a state implemented a state-wide AFDC waiver or TANF. Most states implemented waivers in 1995 or 1996; only 11 states implemented a waiver before 1995. Some states obtained AFDC waivers before 1996 but never implemented them. All states but California implemented TANF in 1996 or 1997. A state is defined as a waiver state if it implemented any state-wide waiver before 1995; otherwise, it is a non-waiver state. In the difference-in-differences analysis of national reform (PROWRA), we use the sample of non-waiver states. The assumption is that there is a time lag between the implementation of a waiver policy and its impact on welfare recipients and hence their children; therefore, the 1996 test scores of students in the non-waiver states are unlikely to be affected by any AFDC waivers.

3.3 Time-Varying State Characteristics

To account for variation in measurable educational inputs, we control for a vector of time-varying state characteristics: pupil-teacher ratio, expenditure per student, 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. The first two reflect measurable school inputs, and the remaining variables reflect forces that may affect family inputs. For 4^{th} graders, 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. All dollar values are in constant 1983 dollars. Table 3 reports summary statistics of these 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 also increased overall.

3.4 Introduction of State School Accountability Systems

States began introducing school accountability systems in the early 1990s. Their effect on academic performance is unlikely to be captured by measurable school input variables. We are able to isolate the effects of welfare reform on school performance from that of school accountability systems because of the uneven introduction of systems over time.

Following Hanushek and Raymond (2005), an accountability system is defined as publishing outcome performance on standardized tests for each school along with providing a way to aggregate and interpret the school performance. 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 Table 2 report the year each state introduced either type of accountability system, ranging from 1993 to 2003.[12] Table 4 summarizes the pattern of introduction of the accountability system; by 2003, 19 states and the District of Columbia had only the report-card accountability.

The No Child Left Behind Act of 2001 (NCLB) was passed in January of 2002. It required stronger accountability of schools in all states, and that all states complete and submit to the Department of Education by January 31, 2003, the Consolidated State Application Accountability Workbook, detailing the state plan for implementing its accountability system under NCLB. In reality, all states submitted their Workbooks between January 2003 and June 2003, and all the state plans were approved by June 2003.[13] Hence, we consider 2003 the year the 20 report-card states introduced the consequential accountability system. We define a dummy variable equal to 1 if an accountability system was introduced prior to the year of a NAEP test. We also use information on the timing of accountability introduction to calculate the years elapsed since a state implemented an accountability system.

4. Main Estimation Results: Fourth Grade Math Scores

4.1 Effects of National Welfare Reform on Fourth Grade Math Scores

We first estimate Equation (2) using fourth-grade math score as the performance measure. For states without state-wide waivers prior to 1995, we compare changes in scores following TANF between low-income and higher-income students. The results in Table 5 show a positive effect of welfare reform on 4th grade math scores, which improved for the FLE students relative to the in-

[12] Hanushek and Raymond (2005) report that Connecticut, North Carolina, and Wisconsin introduced a consequential accountability in "2003 or earlier".

[13] The timing of submission and approval is noted in the state files available at http://www.ed.gov/admins/lead/account/stateplans03/index.html.

eligible students in the decade following national welfare reform. This is clear from the $\beta_{FLE,t}$ terms ($\beta_{FLE,2000}$, $\beta_{FLE,2003}$, $\beta_{FLE,2005}$), which are consistently positive across years and specifications.[14]

The first column reports results from the estimation of Equation (2) with no covariates (X_{ist} contains only ones), a pure DD. 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 positive interaction terms, indicating larger gains for free lunch eligible students, are increasing over time and significantly different from zero (at the 1% level) for 2003 and 2005. The other coefficients in that column indicate that males outperformed females by 1.3 points, whites outperformed non-whites by 15.4 points, and free lunch eligible students lagged their peers by 18 points. The significant year effects show that 4th grade math scores were improving overall during the period.

Consistency of the estimates requires that test score trends for FLE and FLI would have been identical absent welfare reform. This assumption is likely too strong, considering that the period under question experienced dramatic changes in macro-economic conditions and school inputs. Fortunately, we can observe key variables that may drive changes in test scores, and control for them.

Results with additional controls are presented in the remaining columns. Column 2 adds state-year level controls for adult population shares with high school diplomas and college degrees, per capita income (in tens of thousands of dollars), unemployment rate, school expenditures per pupil (also in tens of thousands of dollars), and pupil-teacher ratio. 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.

Next, we address the concern that contemporaneous educational reform aimed at increasing school accountability may have influenced achievement differentially for the two income

[14] For robustness, estimation for each of the empirical approaches was repeated using a logarithmic transformation of the dependent variable, yielding results that were qualitatively identical in sign and significance, but with larger magnitudes for the treatment effects.

groups (Hanushek and Raymond 2005), and the plausible changes in school incentives would not be captured by the controls for spending or pupil-teacher ratios. We estimate Equation (2) including an indicator for the presence of consequential accountability programs, a linear measure of years since the accountability program was initiated, and interactions of FLE with each of these.[15] This is our preferred basic empirical specification. Results are reported in Column 3 of Table 5. Accountability has a positive and significant effect on test scores, and years since accountability has a positive but insignificant effect. The FLE interactions are both negative and insignificant. The new coefficients on the free lunch-year interactions are similar to prior estimates, but somewhat larger in magnitude.

In the next column of the table, we add a complete set of fixed effects for state and year interactions. This specification controls for all common observable and unobservable factors that determine test scores for a given state and year. The comparison is the performance gap between FLE and FLI students within a state and a year, and the $\beta_{FLE,t}$ terms capture average changes in these gaps from 1996 to 2000, 2003 and 2005. Naturally, the state-year controls are omitted from the table. The final column contains the full specification with additional interactions between all of the state-year controls from Column 2 with the FLE indicator, thereby allowing them to produce different effects on low-income and higher-income students. None of the interactions is individually significant. In both specifications, our main variables of interest, the $\beta_{FLE,t}$ terms, are again positive, increasing, and significant starting in 2003.[16] The robustness of the main estimates provides support for the validity of the natural experiment framework.

Welfare reform has a nontrivial effect on test scores, irrespective of model specification. For example, consider the preferred basic specification in Column 3 of Table 5. Relative to base-

[15] Results were not at all sensitive to omitting the linear time trend, or to including report-card programs in the measure of school accountability.

[16] Although about the state samples were designed to include approximately the same number of students, the underlying cell sizes vary within states by race, sex and income. To address the potential heteroskedastisity and improve the efficiency of our estimates, we repeated the regressions using weights for state population shares in the NAEP. Further, in order to produce estimates that are representative of the average student in the nation, we used weights based on the US population (putting more weight on cells from larger states). The results were qualitatively unchanged and quantitatively similar. For example, in the full specification (corresponding to column 5 in Table 5), the estimates were FLE*2000 = 0.705 (s.e. 0.792), FLE*2003 = 2.73 (s.e. 0.795), FLE*2005 = 3.173 (s.e. 1.026) using the first set of weights, and FLE*2000 = 1.903 (s.e. 0.939), FLE*2003 = 3.173 (s.e. 0.976), FLE*2005 = 3.234 (s.e. 1.190) using the second.

line trends between 1996 and 2005, the free lunch eligible students experienced 25% greater test score growth. This extra growth is equivalent to 27% of a standard deviation of the initial 1996 test score distribution of the FLE group. To interpret the magnitude of these 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 17% of a standard deviation in kindergarten year. 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. As an outcome the initial 1996 test score gap between FLE and FLI groups was narrowed by 17% in 2005. Furthermore, these estimates may understate the effects on program participants, since welfare eligible children account for under 50% of all FLE students (footnote 7).

4.2 Changes in Student Behavior and Home Inputs

To provide evidence regarding the mechanism generating the relative test score gains documented in the previous section, we examine changes in self-reported behavior and home environment for 4th grade NAEP test-takers. As part of the NAEP instrument, students are asked to complete a survey about their background characteristics. Using the sample of states without state-wide waivers prior to 1995, we compare changes for the FLE and FLI groups, relative to their respective baseline 1996 pre-reform values.

In response to the question "How often do you discuss things you have studied in school with someone at home?" students in both FLE and FLI groups report higher average values in 2005 than in 1996, with a far larger increase for FLE than FLI students. The mean frequency goes from 8.2 to 10.0 discussions per month for FLE, a 22% increase, and from 9.1 to 10.2 for FLI, an 11% increase.[17] 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 dis-

[17] Means were computed by assigning the following numerical values for frequency per 4 week period to each of the categorical response options: 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.

tribution. Both groups exhibit increasing shares of students in the lowest category, but increased representation in the "zero" bin is 4 times larger for FLI than for FLE 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, showing a relative gain for FLE students in the frequency of discussing their studies at home. Despite the lack of information regarding the content or duration of these discussions, this rough measure suggests that welfare reform was associated with greater parental investment in child education in low-income households.

The survey contains another measure that sheds light on the actual change in time inputs used for building human capital: the amount of time devoted to television and video watching on school days. In 1996, over a quarter of FLE students and more than one eighth of FLI 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 FLE children (a decline of 35.4 minutes per day, or 15%), and from 3.15 to 2.91 for FLI children (a 14.7 minute, or 7%, decline).[18] The relative decrease in television viewing for FLE students is consistent with a relative increase in FLE student investment in school-work and other human-capital investment activities.[19]

The changes in child behavior associated with welfare reform may be the result of improved parental disciplining and role modeling or enhanced stability and daily routine in low-income households.

4.3 Test Score Trends Following Welfare Reform

The second estimation strategy uses data from all states, and exploits variation in the timing of welfare reform stemming from either the adoption of pre-TANF state-wide reform waivers, or from differences across states in the actual implementation date of the national reform. Since our sample period only begins in 1996, we are unable to repeat the exercise of Table 5 to estimate

[18] Means were computed by assigning the following numerical values to each of the 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.

[19] Although it would be useful to corroborate this with reported changes in time spent on studies, the question about homework was only asked in 1996 and 2000. Responses to a question about time per day devoted to math homework were very stable between 1996 and 2003 for both income groups.

the effects of pre-TANF waivers using before and after observations. Instead we estimate the differential relationship between test scores and time elapsed since reform between low-income and higher-income students. Due to the limited number of years in our dataset, we also confine our attention to a linear measure of time since welfare reform. If the true function is non-linear, the linear term will capture the average slope during the period. In the regression equation:

$$P_{ist} = \beta_s + \beta_{FLE} \cdot FLE_i + \beta_t \cdot \tau_t + \beta_{FLEMSR} \cdot FLE_i \cdot MSR_{st} + \beta_{MSR} \cdot MSR_{st} + \beta_X \cdot X_{ist} + \nu_{ist}, \quad (3)$$

our variable of interest is $FLE_i \cdot MSR_{st}$, the interaction term between MSR_{st} (months since reform) and free lunch eligibility. As shown in Table 6, the parameter estimate is positive and statistically significant across the various specifications, confirming the prior finding of a positive effect of reform on achievement.

Column 1 contains results from the parsimonious specification, with controls for race, sex, free-lunch, and state and year fixed effects. The estimate for β_{FLEMSR} of 0.020 (with a robust standard error of 0.005) is significant at the 1% level. This implies that the test score gains of free lunch eligible students relative to ineligible students grow as months pass following welfare reform. The inclusion of controls for state-year income, unemployment, education levels and educational spending and staffing (Column 2) and educational reform (Column 3) does not alter the estimate. Finally, the last two columns include a full set of state and year fixed effect interaction terms, and the last also allows for the state-year level controls to vary in their effects by free lunch status. The point estimate increases in Column 4 to 0.027, and decreases in Column 5 to 0.018, and remains significantly different from zero at the 5% level.

The final column provides a robustness check to address a new concern that arises for this method – the potential endogeneity of the timing of state-level reforms. The estimates of β_{FLEMSR} in the previous columns are identified from variation both within states over time, and across states, based on their timing of reform implementation. The FLE fixed effect included in the regressions absorbs national differences in scores between FLE and FLI students, but does not address the potential correlation between timing of adoption and the initial test score gap between low-income students and others. In particular, if states with smaller test score gaps adopted earlier, β_{FLEMSR} will be biased upwards. In column 6, we address differences in levels with an additional set of fixed effects, interactions between FLE_i and state indicators. These

terms remove any average differences in test score gaps across states. The new estimate of β_{FLEMSR} is 0.041 with a standard error 0.015, showing that the effect is robust to controlling for differences in levels.[20] This value is roughly comparable with the previous estimates. To see this, we can take the measure of the total effect over the period, captured in the free lunch-year 2005 interaction in Table 5, and divide it by 108 for the months passed since 1996. The range of estimates is from 0.024 to 0.030.

As in the previous tables, we find that whites score higher than non-whites, boys score higher than girls, and free-lunch eligible students score lower than their counterparts from wealthier households. There is a positive (significant) and growing (not significant) benefit from accountability on math scores.[21] The β_{MSR} terms are statistically insignificant, indicating that welfare reform has indeed no effect on free-lunch ineligible students.

4.4 Test Score Trends Following Implementation of Different Sanction Rules

Although the above method accounts for variation in the timing of implementation of welfare reform across states, it still groups together states with a wide range of welfare policy characteristics. States' welfare policies varied in their generosity and eligibility rules under AFDC, and TANF introduced far more variation. While we might be interested in exploring which aspects of welfare are most salient, we face the problem of all researchers interested in specific welfare components: there are too many dimensions along which policies can vary and too much correlation in choices by states to isolate the different features (Blank 2002). In addition, since the policy variables are choices, we may be concerned with endogeneity, more so than for the timing decision. Although a complete analysis of the different components is infeasible, we can explore a narrow, well-defined, and theoretically important dimension – the penalties imposed on welfare recipients for violations of their work requirements. Under the AFDC waivers and TANF, individuals are required to look for jobs or to work after a certain period of being on the welfare roll; violation of these work requirements results in either a reduction or a complete loss of the welfare benefit. The sanction rules appear to provide a direct and immediate incentive for single mothers on welfare to engage in labor market activities.

[20] Unfortunately, since the first observation in the data is 1996, it is impossible to assess pre-existing trends.
[21] Here again, our preferred specifications are those with controls for accountability reform, found in columns 4 to 6 of the table. We report results from the other specifications for completeness and comparability.

We define four types of sanctions based on rules on the reduction in cash benefit the first time a welfare recipient violates the work requirements (initial sanction) and on the most severe sanction for violating the work requirements. The omitted group is "no sanction." "Partial sanction" applies if welfare benefit is reduced partially in both the initial sanction and the most severe sanction; "gradual sanction" applies if welfare benefit is reduced partially in the initial sanction and is reduced fully in the most severe sanction; "full sanction" applies if welfare benefit is reduced fully in both the initial sanction and the most severe sanction. Table 7 reports the time a state first started any type of sanction, the times a state switched to a gradual or full sanction if it ever did, and the types of sanctions at these times. States started to implement any type of sanction policy either under the AFDC waiver or the TANF. At the commencement, 29 states had a partial sanction; half the remaining had a gradual sanction, and the other half a full sanction. Over time, 18 of the initial partial-sanction states switched to gradual sanction; 1 of the initial partial-sanction states (Hawaii) switched to full sanction; and 2 of the initial gradual-sanction states (Michigan and Wyoming) switched to full sanction. No state switched from a more severe sanction to a less severe one.

We construct a new set of variables: $MSS1_{st}$ = months since any sanctions, $MSS2_{st}$ = months since gradual or full sanctions, and $MSS3_{st}$ = months since full sanctions. These are interacted with FLE_i to produce the new variables of interest, in increasing order of severity: $FLE_i \cdot MSS1_{st}$, $FLE_i \cdot MSS2_{st}$, $FLE_i \cdot MSS3_{st}$. The estimation equation is comparable to the previous estimation equation, although instead of a single term (β_{FLEMSR}) for welfare reform, there are three ($\beta_{FLEMSS1}$, $\beta_{FLEMSS2}$, $\beta_{FLEMSS3}$), allowing for heterogeneous effects by sanction severity. Estimates are presented in Table 8, which mirrors the structure of Table 6, with the first column containing the most parsimonious specification. The next column adds the state-year level controls for income, education and school inputs. Column 3 adds the school accountability variables and their interactions with FLE_i, and Column 4 add the state-year fixed effects. Finally, Column 5 adds the interactions between FLE_i and the state-year level controls. The estimates for $\beta_{FLEMSS1}$ are consistently positive and significant at the 10% level, across specifications, showing that even partial sanctions produced a positive effect on test scores. The estimates for $\beta_{FLEMSS2}$ and $\beta_{FLEMSS3}$ are positive but insignificant, and F-tests on $\beta_{FLEMSS2} + \beta_{FLEMSS3}$ failed to reject zero.

This implies that states with stricter penalties for violations of work requirements had test score gains which were statistically indistinguishable from those with partial sanctions.[22]

4.5 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 (FLE) likely contains past, current, and potential welfare eligible children in all years, while the higher-income group (FLI) likely excludes them. Therefore, we argue that children leaving welfare as an outcome of reform are unlikely to switch status from FLE to FLI; 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 (FLE), ineligible (FLI), 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-estimate or under-estimate the true treatment effect.

The size and direction of the bias in our previous estimates of the treatment effect will depend on the performance distributions of students in the three income groups. In what follows we show that rates of non-reporting declined over the period. We then provide a new set of *lower-bound* estimates for the treatment effect of welfare reform on math achievement. These estimates are also positive, confirming the beneficial impact of Sections 4.1- 4.4.[23]

[22] Since sanctions failed to provide additional information, we do not report results from these specifications in the robustness checks or heterogeneous effects models that follow. Results are available upon request from the authors.
[23] In the Appendix we discuss in great depth implications of changes in income-reporting for our DD estimator and the validity of our lower-bound estimate.

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 FLE and FLI 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 FLE group from 64% to 71%, a negligible increase in FLI from 26% to 27%, and a dramatic decrease in NA from 10% to 2%, over the decade.[24]

The steadily increasing trend in reported free-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,[25] suggesting that the apparent 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 FLE or FLI 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 FLE shares, as they will not generally lead to a relaxation of standards.

To test the robustness of our estimates, 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 FLE groups to generate the average performance of the low-income group. This counter-factual experiment produces our *lower-bound* estimates of welfare reform's effect. Table 9 shows the estimates produced by assigning the entire NA group to the FLE 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 3, $\beta_{FLE,2000} = 0.512$, $\beta_{FLE,2003} = 1.487$ (significant at 5%), and $\beta_{FLE,2005} = 1.328$. Hence, even our *lower-bound* point estimates are positive, although not always statistically significant. The counterfactual estimates

[24] Year 2000 data for non-whites are not included in the table because reporting standards were not met.
[25] Source for poverty rate data is http://www.census.gov/hhes/www/poverty/poverty.html .

for β_{FLEMSR} are smaller than the original estimates (found in Table 6), 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.

5 Heterogeneous Effects of Welfare Reform on Test Scores

There is no reason to assume that the effects of welfare reform are uniform across all children. Rather, the test score response may vary according to observable and unobservable characteristics of the treated population. This section evaluates the heterogeneity of the treatment effect by student ability, sex, race, and grade level.

5.1 Test Score Distribution

The previous results are estimated using cell average scores as the dependent variable. This method can hide important variation for students with different ability levels (within their own income group), and may produce estimates that suffer from attenuation bias if there is substantial clumping of scores at either of the extremes (0 or 500).

In Table 10, we present estimates separately for the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the score distribution. Panel A shows results with the basic set of controls. Point estimates at the median are very close to earlier estimates for means. For the DD estimates from the first method, there is no consistent pattern of treatment effects across the ability distribution. For example, $\beta_{FLE,2000}$ and $\beta_{FLE,2005}$ increase for higher ability students, but $\beta_{FLE,2003}$ decreases and increases. The β_{FLEMSR} coefficient tends to be higher for lower ability levels. Panel B shows results with the full set of controls. Positive and significant effects of welfare reform are found across the entire score distribution.

That higher-ability, low-income students also benefit from welfare reform indicates that the positive treatment effects found in previous sections are not an artifact of mean regression, in which low-performers have larger gains because of their lower starting point. The overall smaller estimates for higher-ability students also suggest that there might be slight diminishing returns to improved family environments for this group of students.

In an additional robustness exercise, we repeated the regressions by test score percentile, but instead of matching the 10th percentile in the FLI group with the 10th percentile in the FLE group, we matched it with the 25th percentile, creating a control group that had similar achievement to the treatment group in the period before welfare reform. Similarly, the 25th percentile FLI was matched with the median FLE, the median FLI with the 75th FLE, and the 75th FLI with the 90th FLE. As shown in Panel B of Table 1, these new grouping produce very similar initial 1996 test scores between the FLE and FLI students, a major advantage to the technique, which allows us to avoids the critique of that the quasi-experiment is invalid due to a lack of comparability between the treatment and control groups.

The results of the exercise confirm the prior findings, and are reported in Appendix Table 1. As above, welfare reform is associated with a relative improvement in test scores for economically disadvantaged children, generally statistically significant in 2003 and 2005, and somewhat larger for children at the lower end of the test score distribution. This confirms that the apparent gains are neither the result of mean reversion nor are they driven by uncontrolled factors that operate through initial test score differences between the groups.

5.2 Sex and Race

We might expect welfare reform to have larger, more positive effects on the test scores of girls relative to boys for two main reasons. First, if there is a positive effect of maternal employment on child achievement through positive role model effects, and if children are affected more by their same-sex parents, we should expect these effects to be larger for daughters than for sons. The differential role model effect that favors daughters may be lessened if sons and daughters in single-parent households are equally influenced by their mother's activities in setting their own job and life aspirations. Second, the diminished attractiveness and feasibility of welfare as a permanent substitute for employment may have a larger impact on the human capital investment of girls than boys, since women are more likely to be eligible for welfare than men. Alternatively, we may expect larger effects for boys if welfare reform changed family structure in such a way that boys became more likely to have contact with their fathers. Also, larger effects for boys

may result mechanically from the fact that boys have higher average scores and greater score variance than girls.

Results by sex are reported in Panel A of Table 11. For each dependent variable and subgroup, two specifications are reported, corresponding to the preferred basic model (Column 3 in Tables 5 and 6) and the full model (Column 5 in Tables 5 and 6). Estimates are reported for the regressors of primary interest in the first and second empirical approaches. With the basic set of controls, the $\beta_{FLE,t}$ terms are larger for boys than for girls, although these differences are not statistically significant, and the β_{FLEMSR} estimates are indistinguishable. Under the full set of controls, the $\beta_{FLE,t}$ and β_{FLEMSR} estimates are larger for girls, but not statistically so. Hence, there is no consistent evidence of heterogeneity by sex.

In the case of race, we may expect larger effects for Non-Whites, because they are more likely to qualify for welfare, conditional on being eligible for the federal subsidized or free lunches. Results by race are reported in Panel B of Table 11, and again there is 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.

5.3 Grade Level

We repeat the DD analysis using a second measure of academic performance, 8th grade math scores. We again use the NAEP test scores for 1996, 2000, 2003, and 2005. Therefore, the birth cohorts of the 8th graders are 1983, 1987, 1990, and 1992. For the early years, the difference in the treatment effect between 4th and 8th grade scores is essentially due to the difference in age at first exposure to welfare reform. Heterogeneous effects by age at first exposure may provide additional information about the relative importance of the two primary channels for the net effect of welfare reform: changes in parental inputs and direct changes in child expectations. Finally, in this section and the next, we show that test score gains found for fourth grade students had some permanence, and that their scores continued to improve at higher grades.

Table 12 presents estimates from both the DD model (Equation (2)) and from the analysis of time trend following welfare reform (Equation (3)). The new dependent variable is average 8th

grade test score. The first two columns report the estimates of Equation (2) from the preferred basic specification and the full specification, and the last two columns report the estimates of Equation (3) from the two specifications.[26]

Noticeable differences exist between Table 5 and Table 12. While males continue to outperform females, the size of the gap is much smaller, about 0.6 points. The score advantage for whites relative to non-whites is larger in the eighth grade, at about 20 points. The free lunch disadvantage is similar for 4th and 8th grades, at about 17-18 points. The overall trend in scores is again increasing relative to 1996, but not all of the year effects are statistically significant. The state-year level controls are generally insignificant, as are their interactions with *FLE*. As with 4th grade, the coefficients on strong school accountability reform and the trend following accountability reform are both positive; unlike the 4th grade coefficients, however, these coefficients are statistically significant at the 1% and 5% levels, respectively. In an important departure from the 4th grade results, the interactions between free lunch and accountability are statistically significant at the 5% level. While the coefficient for 4th grade scores is also negative, it is not significant. Thus, the inclusion of the accountability terms, and their interactions with free lunch, is even more important for 8th grade scores than for 4th grade.

In comparison with the 4th grade results, the estimates of $\beta_{FLE,t}$ are smaller and less significant. In the basic specification, the estimate for 2000 is negative, although not significant, and the estimates are insignificant for both 2003 and 2005. In the full model, the estimates for 2000 and 2003 are smaller than their counterparts in the 4th grade regression and are insignificant, but the estimate for 2005 (3.282) is significant and of comparable magnitude as that in the 4th grade regression. The point estimates for β_{FLEMSR} in the preferred basic model and full model are 0.003 and 0.007 respectively. They are far smaller than the estimates of 0.020 and 0.041 for 4th grade, and are not significantly different from zero.

One implication of the table is that welfare reform had a more positive effect on fourth grade students than eighth grade students, conditional on the same duration of exposure to re-

[26] Of the control variables for 8th graders, school inputs are averages over 1988-1995, 1992-1999, 1995-2002, and 1997-2004 periods; other variables are averages over the 1983-1995, 1987-1999, 1990-2002, and 1992-2004 periods for the test years.

form. This suggests that, separate from duration of exposure, the beneficial effects of welfare reform are stronger for children exposed to the reform at an earlier age. Even within the four year age difference between fourth graders (9 year olds) and eighth graders (13 year olds), substantial differences are apparent. The fact that the effects are stronger for younger children also indicates that family factors, changes in maternal behavior that influence child human capital investment, are an important mechanism for the reduced form effect, possibly more important than direct student responses to welfare reform. Reform made welfare receipt a less attractive future option for young people; if the effect of reform were driven only by direct student response to this fact, we should expect larger effects for older children.

Another implication of the table is that the benefits of welfare reform on fourth grade achievement appear to have some permanence and to be reflected in eighth grade scores for those same students four years later. We compare the DD estimates for 4th grade students in 2000 with the average of the coefficients for 8th grade students in 2003 and 2005. The coefficients for 4th graders from the preferred basic and full models are 0.880 and 0.674, in columns 3 and 5 respectively. The average values for 8th graders, from the same specifications, are 1.114 and 2.385. Hence, it appears that the fourth grade gains carried over to the eighth grade, and may have even increased.[27]

6. Value-Added Estimation

We are concerned that the relative test score gains for free-lunch eligible students reported in the previous sections may in part be attributable to major policy changes that occurred just a few years *before* PRWORA and also affected the income and work incentives of low-income families: expansions of the earned income tax credit (EITC). The most recent expansions of the EITC occurred in 1993, and preceded both national welfare reform and the start of our sample period. 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 im-

[27] Note that this cohort was 5 years old when first exposed to welfare reform in 1996. Hence, their gains in academic achievement could not have been an outcome of changes in their early childhood development, caused by expansions to childcare funding. Rather, these results support the conjecture that improved family environments when children grow up are conducive to school performance.

provements in children's academic achievement (Dahl and Lochner 2005). Therefore, EITC expansions are expected to have similar effects on low-income families as welfare reform; in turn, their effects on school performance will be of the same direction as that of the welfare reform. Not controlling for EITC will bias the estimated effect of welfare reform upward. The timing of the EITC expansions implies that their effects on school performance will mostly work through their effects on family inputs during a test-cohort's early childhood.

Without measured achievement data for earlier years, direct estimation of EITC's contribution to the achievement gap is impossible. Instead, we provide supportive evidence to our baseline estimation of Equation (2) by estimating it in a value-added framework. In so doing, we make use of the fact that the 1987 cohort was tested in 1996 as 4th graders and in 2000 as 8th graders, and that the 1991 cohort was tested in 2000 as 4th graders and could potentially be tested in 2004 as 8th graders. 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 remove the effect of any confounding factors that take effects during the early childhood and during the first four years of schooling. The idea is illustrated as follows:

$$P_{ijct}^{g=8} = f(P_{ijc,t-4}^{g=4}, F1_{ijc}^{(t-4,t)}, 1(j=1)\cdot F2_{ijc}^{(t-4,t)}, S_{ijc}^{(t-4,t)}) + \nu_{ijct} - \nu_{ijc,t-4}, \qquad (1'')$$

where $c = 1987$ and 1991, $t = 2000$ and 2004.[28] 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):

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

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

[28] 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.

Partial estimation results of Equation (4) from our full specification are reported as follows. Numbers in parentheses are robust standard errors clustered at state level.

$$P_{ist}^{g=8} = P_{is,t-4}^{g=4} \cdot 1.07 + 3.98 \cdot 1(FLE_i = 1) \cdot 1(t = 2004) + \ldots\ldots + \hat{\varepsilon}_{ist} .$$

(0.03) (1.67)

The control for 4th-grade score removes differences in initial stock of human capital, and its coefficient is close to unity. The value-added estimate of welfare reform's effect is captured by the coefficient estimate, 3.98, on the interaction between the FLE indicator and year 2004. This positive and significant (at the 5% level) estimate indicates that EITC cannot explain the change in the gap in test scores between free-lunch eligible students and ineligible students, and that welfare reform has played a role. Relative to welfare reform's effect on the growth of math score between 1996 and 2000, its effect on the growth of math score between 2000 and 2004 is 3.98 points. Coefficient estimates on other control variables are generally consistent with those in Table 12. When we examine male and female, white and non-white separately, the point estimates range from 3.13 to 4.81.

Although the exercise provides additional support for the claim that welfare reform was indeed the cause of the measured test score gains for low-income students, data limitations prevent a more conclusive strategy. If one prefers to remain agnostic regarding the role of EITC in producing the gains measured in the previous sections, the basic structural implications nevertheless remain unchanged. Instead of attributing the gains to welfare reform alone, the gains are from the combined effects of EITC and welfare reform. The key labor supply effects of EITC are qualitatively identical to those of welfare reform: increasing work hours and participation of low-income parents. Although the response intensity and distribution across demographic groups may differ between the two policies, the results from either policy imply that increased parental work hours help students in low-income families.

7. Pre-Existing Trends in the Achievement Gap

The empirical results in this paper provide evidence that welfare reform was associated with a narrowing of the achievement gap between low-income and higher-income students in the

United States. We supplement the main DD framework with estimates that exploit variation in the timing of waiver adoption and TANF implementation, and use a value-added framework to remove the potential effects of the EITC expansions. However, one may still be concerned that observed narrowing during our sample period was not caused by welfare reform, but rather the result of a pre-existing trend of a narrowing achievement gap by income. It is not possible to address this concern directly using the NAEP because free lunch eligibility is only available starting in 1996. Nevertheless, Figure 4 provides suggestive evidence against a pre-existing narrowing by plotting the changes in national average test scores, relative to 1996 baseline, by race and ethnicity.[29] Eligibility for the free lunch program is highly correlated with race, and both African Americans and Latinos are more likely than Whites to be eligible.[30] In the years preceding welfare reform, test scores for white students improved relative to those for Latino students and were stable relative to those for African American students. However, after 1996, and especially after 2000, African Americans and Latinos showed relative test score gains. The time pattern is consistent with a break from trend. Allowing for the limitations inherent in the exercise, it is useful to note that the results support a causal interpretation for the main findings of the paper.

8. 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 of academic achievement that spans a ten year period starting just prior to the welfare reform. We perform reduced-form estimation of the net effect of welfare reform. In a difference-in-differences framework, we use high-income group as the control group to simulate the unobservable counterfactual time trend, and use control variables for the economic and educational environment, and additional fixed effects and interaction terms. There is no evidence that welfare reform harmed the academic performance of poor students. On the contrary, the key finding is improvement in the relative math scores of fourth-grade students from low-income families, and the improvement is statistically significant for years 2003 and 2005. There is also improvement

[29] 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.

[30] FLE students constitute 23% of Whites, 72% of African Americans and 69% of Latinos.

in the relative math scores of eighth-grade students from low-income families, although it is of smaller magnitude and is significant only in 2005. These improvements are present for all race, gender and ability groups. The results for 4th and 8th graders jointly suggest that the beneficial effect of early exposure to the welfare reform does not disappear later; if anything, it is strengthened over time. Preliminary analysis suggests that these beneficial effects are related to relative improvements in low-income children's time use and interaction with parents. A thorough exploration of these and other channels for the underlying mechanism of the test score gains is 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.

Chase-Lansdale, P.L., R.A. Moffitt, B.J. Lohman, A.J. Cherlin, R.L. Coley, L.D. Pittman, J. Roff, and E. Votruba-Drzal, *Mothers' Transitions from Welfare to Work and the Well-Being of Preschoolers and Adolescents.* Science, 2003. 299(5612): p. 1548-1552.

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.

Fender, L., S.M. McKernan and J. Bernstein, *Linking State TANF and Related Policies to Outcomes: Preliminary Typologies and Analysis*. The Urban Institute, 2002. Available at http://minerva.urban.org/typologies/.

Figlio, D.N. and J.P. Ziliak, *Welfare Reform, the Business Cycle and the Decline in AFDC Caseloads*, in *Economic conditions and welfare reform*, S.H.e. Danziger, Editor. 1999.

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.

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.

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.

Haveman, R. and B. Wolfe, *The Determination of Children's Attainments: A Review of Methods and Findings*. Journal of Economic Literature, 1995. 33(4): p. 1829-1878.

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 Economimcs, 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.

Moffitt, R., *Economic Effects of Means-Tested Transfers in the U.S.*. National Bureau of Economic Research, Inc. 2002, NBER Working Papers: w8730.

Moffit, R.A., *The Temporary Assistance for Needy Families Program*. In Robert A. Moffit ed *Means-Tested Transfer Programs in the United States,* the University of Chicago Press, 2003.

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.

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.*

Rowe, G., *The Welfare Rules Databook: State Policies as of 1999*. The Urban Institute, 2000. Available at http://www.urban.org/UploadedPDF/wrd.pdf.

Rowe, G. and V. Russell, *The Welfare Rules Databook: State Policies as of 2002*. The Urban Institute, 2004. http://www.urban.org/UploadedPDF/311109_DP04-06.pdf.

Rowe, G. and J. Versteeg, *The Welfare Rules Databook: State Policies as of 2003*. The Urban Institute, 2005. http://www.urban.org/UploadedPDF/411183_WRD_2003.pdf.

Ruhm, C.J., *Parental Employment and Child Cognitive Development.* Journal of Human Resources, 2004. 39(1): p. 155-92.

Schoeni, R.F. and R.M. Blank, *What has Welfare Reform Accomplished? Impacts on Welfare Participation, Employment, Income, Poverty, and Family Structure*. National Bureau of Economic Research, Inc. 2000, NBER Working Paper; w7627.

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 dynamics: Theory and evidence*. Working Paper, Department of Economics, Clemson University, 2005.

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: 4th grade Math Score Summary Statistics

Panel A: By School-lunch Eligibility

Year		*All*	*FLI*	*FLE*
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

		FLI					*FLE*				
year		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

		FLI				*FLE*			
Year		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

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

Table 2: Welfare Reform and School Accountability Dates

State	Earliest State-Wide Waiver Implemented	TANF Implemented	Consequential Accountability Introduced	Report-Card Accountability Introduced
Alabama		November-96	1997	
Alaska		July-97		2001
Arizona	November-95	October-96		2000
Arkansas	July-94	July-97	1999	
California	December-92	January-98	1999	
Colorado		July-97		2002
Connecticut	January-96	October-96	1993	
Delaware	October-95	March-97	1998	
DC		March-97		1997
Florida		October-96	1999	
Georgia	January-94	January-97	2000	
Hawaii	February-97	July-97		2001
Idaho		July-97		1997
Illinois	November-93	July-97		1999
Indiana	May-95	October-96		1995
Iowa	October-93	January-97	2003	
Kansas		October-96		1995
Kentucky		October-96	1995	
Louisiana		January-97	1999	
Maine		November-96		1999
Maryland	March-96	December-96	1999	
Massachusetts	November-95	September-96	1998	
Michigan	October-92	September-96	1998	
Minnesota		July-97		1996
Mississippi	October-95	July-97		1994
Missouri	June-95	December-96		1997
Montana	February-96	February-97		1998
Nebraska	October-95	December-96		2001
Nevada		December-96	1996	
New Hampshire		October-96		1993
New Jersey	October-92	July-97	1997	
New Mexico		July-97	2003	
New York		November-97	1998	
North Carolina	July-96	January-97	1993	
North Dakota		July-97	2003	
Ohio	July-96	October-96	1997	

State	Earliest State-Wide Waiver Implemented	TANF Implemented	Consequential Accountability Introduced	Report-Card Accountability Introduced
Oklahoma		October-96	1996	
Oregon	February-93	October-96	2000	
Pennsylvania		March-97		1999
Rhode Island		May-97	1997	
South Carolina		October-96	1999	
South Dakota	June-94	December-96		1997
Tennessee	September-96	October-96	1996	
Texas	June-96	November-96	1994	
Utah	January-93	October-96	2003	
Vermont	July-94	September-96	1999	
Virginia	July-95	February-97	1998	
Washington	January-96	January-97		1998
West Virginia	February-96	January-97	1997	
Wisconsin	January-96	September-97	1993	
Wyoming		January-97		1999

Sources: Crouse (1999) for the dates of state waiver implementation and TANF implementation. 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.

Table 3: Control Variable Summary Statistics

Year		*Pupil-Teacher Ratio (K-12)*	*Per Pupil Expenditure, (K-12, $)*	*Population % HS Diploma (age ≥ 25)*	*Population % College Degree (age ≥ 25)*	*Average Income ($)*	*Unemploy-ment Rate*
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

Sources: 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.

Table 4: Number of States Introducing School Accountability

Year	Consequential Accountability	Report Card Accountability
1993	3	1
1994	1	1
1995	1	2
1996	3	1
1997	5	4
1998	5	2
1999	7	4
2000	2	1
2001	0	3
2002	0	1
2003	4	0
Total	31	20

Note: Counted from the last two columns of Table 2.

Table 5: National Welfare Reform and 4th Grade Math Scores

	1	2	3	4	5
FLE*Year = 2000	0.64	0.665	0.880	1.121	0.674
	[0.725]	[0.723]	[0.709]	[0.723]	[0.769]
FLE*Year = 2003	2.196	2.184	2.474	2.626	2.486
	[0.443]**	[0.450]**	[0.562]**	[0.565]**	[0.736]**
FLE*Year = 2005	2.569	2.57	3.062	3.234	3.041
	[0.522]**	[0.539]**	[0.881]**	[0.884]**	[0.916]**
Year = 2000	3.815	3.781	2.828		
	[0.707]**	[1.095]**	[1.021]**		
Year = 2003	13.103	12.84	12.002		
	[0.815]**	[1.698]**	[1.405]**		
Year = 2005	16.467	15.367	13.225		
	[0.830]**	[2.190]**	[1.863]**		
White	15.066	15.069	15.082	15.009	15.06
	[0.535]**	[0.534]**	[0.531]**	[0.558]**	[0.550]**
Male	1.329	1.333	1.336	1.337	1.341
	[0.248]**	[0.248]**	[0.249]**	[0.264]**	[0.264]**
FLE	-17.957	-17.946	-17.91	-18.013	-16.139
	[0.503]**	[0.510]**	[0.517]**	[0.558]**	[6.063]*
Pupil-Teacher Ratio		1.188	1.20		
		[0.511]*	[0.557]*		
Accountability			2.282		
			[0.756]**		
Years of Accountability			0.205		
			[0.189]		
FLE*Accountability			-0.218	-0.411	-0.156
			[0.783]	[0.831]	[0.669]
FLE*YrsAccount			-0.061	-0.043	0.043
			[0.103]	[0.108]	[0.118]
Constant	209.848	163.779	163.415	209.381	154.516
	[0.664]**	[23.634]**	[22.656]**	[0.446]**	[2.800]**
Observations	1034	1034	1034	1034	1034
R-squared	0.94	0.94	0.94	0.96	0.96

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 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 controls for state accountability system; Column 4 adds state*year fixed effects; Column 5 adds interactions between *FLE* indicator and time-varying controls. Coefficient estimates not reported for all included control variables.

Table 6: Welfare Reform Timing and 4th Grade Math Scores

	1	2	3	4	5	6
FLE*Months Since Reform	0.020	0.020	0.020	0.021	0.018	0.041
	[0.005]**	[0.005]**	[0.008]*	[0.008]*	[0.008]*	[0.015]**
Months Since Reform	0.035	0.057	0.032			
	[0.076]	[0.072]	[0.061]			
White	15.129	15.131	15.119	15.044	15.083	15.102
	[0.480]**	[0.481]**	[0.480]**	[0.509]**	[0.503]**	[0.514]**
Male	1.474	1.474	1.478	1.474	1.477	1.477
	[0.218]**	[0.219]**	[0.220]**	[0.233]**	[0.234]**	[0.239]**
FLE	-18.014	-18.003	-18.039			
	[0.503]**	[0.510]**	[0.496]**			
Accountability			1.635			
			[0.671]*			
Years of Accountability			0.279			
			[0.190]			
FLE*Accountability			0.147	0.016	0.192	0.166
			[0.684]	[0.723]	[0.650]	[0.549]
FLE*YrsAccount			-0.021	-0.014	0.142	-0.033
			[0.102]	[0.108]	[0.105]	[0.131]
FLE*HS Graduates					0.243	0.126
					[0.086]**	[0.172]
FLE*College Graduates					-0.033	-0.386
					[0.114]	[0.323]
FLE*Income					-7.104	0.797
					[3.035]*	[7.320]
Constant	209.847	194.966	199.508	209.324	169.796	172.12
	[0.627]**	[28.943]**	[28.162]**	[0.446]**	[2.771]**	[1.896]**
Observations	1309	1309	1309	1309	1309	1309
R-squared	0.94	0.94	0.94	0.96	0.96	0.96

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 5.

Table 7: Welfare Reform Sanction Dates and Types

State	Initial Sanction Date	Initial Sanction Type	Initial G or F Sanction Date	Sanction Type	Starting date of F type	End Sanction Type
Alabama	November-96	P	May-97	G		G
Alaska	July-97	P	July-02	G		G
Arizona	November-95	P	August-97	G		G
Arkansas	July-97	P	August-01	G		G
California	January-98	P				P
Colorado	July-97	G	July-97	G		G
Connecticut	January-96	G	January-96	G		G
Delaware	March-97	P				P
DC	March-97	P	July-97	G		G
Florida	October-96	F	October-96	F	October-96	F
Georgia	January-94	P	July-97	G		G
Hawaii	July-97	P	August-99	F	August-99	F
Idaho	July-97	F	July-97	F	July-97	F
Illinois	October-95	P	July-97	G		G
Indiana	May-95	P	May-03	G		G
Iowa	October-93	F	October-93	F	October-93	F
Kansas	October-96	F	October-96	F	October-96	F
Kentucky	October-96	P	November-98	G		G
Louisiana	January-97	P	March-98	G		G
Maine	November-96	P				P
Maryland	October-96	F	October-96	F	October-96	F
Massachusetts	November-95	G	November-95	G		G
Michigan	October-94	G	October-94	G	April-97	F
Minnesota	July-97	P	July-01	G		G
Mississippi	July-97	F	July-97	F	July-97	F
Missouri	June-95	P				P
Montana	February-97	P	May-98	G		G
Nebraska	December-96	F	December-96	F	December-96	F
Nevada	December-96	P	January-98	G		G
New Hampshire	October-96	P				P
New Jersey	October-92	P	July-97	G		G
New Mexico	July-97	G	July-97	G		G
New York	November-97	P				P
North Carolina	July-96	P	December-99	G		G
North Dakota	July-97	P	July-98	G		G
Ohio	July-96	G	July-96	G		G

State	Initial Sanction Date	Initial Sanction Type	Initial G or F Sanction Date	Sanction Type	Starting date of F type	End Sanction Type
Oklahoma	October-96	F	October-96	F	October-96	F
Oregon	July-96	G	July-96	G		G
Pennsylvania	March-97	G	March-97	G		G
Rhode Island	May-97	P				P
South Carolina	October-96	F	October-96	F	October-96	F
South Dakota	June-94	P	September-97	G		G
Tennessee	September-96	F	September-96	F	September-96	F
Texas	June-96	P				P
Utah	July-96	G	July-96	G		G
Vermont	July-94	P				P
Virginia	February-97	F	February-97	F	February-97	F
Washington	January-97	P				P
West Virginia	February-96	P	January-97	G		G
Wisconsin	January-96	G	January-96	G		G
Wyoming	January-97	G	August-97	F	August-97	F

Sources: Fender et al (2002), Rowe (2000), Rowe and Russell (2004), Rowe and Versteeg (2005), U.S. DHHS (1997), and the Urban Institute Welfare Rules Database
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.

Table 8: Welfare Sanction Rules and 4th Grade Math Scores

	1	2	3	4	5
FLE*MS Any Sanctions	0.017	0.018	0.019	0.021	0.035
	[0.009]+	[0.010]+	[0.010]+	[0.011]+	[0.016]*
FLE*MS G or F Sanctions	0.005	0.006	0.006	0.000	-0.004
	[0.011]	[0.012]	[0.013]	[0.012]	[0.010]
FLE*MS F Sanctions	0.007	0.007	0.006	0.000	0.005
	[0.012]	[0.012]	[0.013]	[0.009]	[0.014]
Months Any Sanctions	-0.106	-0.132			
	[0.068]	[0.057]*			
Months G or F Sanctions	0.000	-0.005			
	[0.018]	[0.017]			
Months F Sanctions	0.022	0.023			
	[0.017]	[0.016]			
White	15.126	15.104	15.043	15.08	15.102
	[0.481]**	[0.481]**	[0.509]**	[0.504]**	[0.514]**
Male	1.475	1.48	1.475	1.478	1.477
	[0.220]**	[0.221]**	[0.233]**	[0.234]**	[0.239]**
FLE	-18.029	-18.05	-18.083		
	[0.484]**	[0.478]**	[0.514]**		
Accountability		1.52			
		[0.618]*			
Years of Accountability		0.398			
		[0.192]*			
FLE*Accountability		0.059	-0.036	0.236	0.251
		[0.676]	[0.713]	[0.645]	[0.569]
FLE*YrsAccount		-0.045	-0.04	0.129	-0.023
		[0.118]	[0.126]	[0.109]	[0.131]
FLE*HS Graduates				0.232	0.139
				[0.088]*	[0.176]
FLE*Income				-7.441	2.295
				[2.975]*	[7.817]
FLE*Pupil-Teacher Ratio				-0.25	1.023
				[0.179]	[0.594]+
Constant	189.567	194.717	209.323	152.589	159.482
	[26.069]**	[23.397]**	[0.441]**	[3.369]**	[2.828]**
Observations	1309	1309	1309	1309	1309
R-squared	0.94	0.94	0.96	0.96	0.96

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 5.

Table 9: Lower-Bound Estimates of Treatment Effects for 4th Grade Math Scores

	1	2	3	4	5
FLE*Year = 2000	0.286	0.348	0.512	0.718	0.133
	[0.787]	[0.791]	[0.805]	[0.816]	[0.839]
FLE*Year = 2003	1.216	1.229	1.478	1.652	1.261
	[0.557]*	[0.561]*	[0.626]*	[0.619]*	[0.758]+
FLE*Year = 2005	0.943	0.967	1.328	1.469	0.887
	[0.662]	[0.675]	[0.899]	[0.886]+	[0.929]
Observations	1054	1054	1054	1054	1054
R-squared	0.93	0.94	0.94	0.95	0.95

	6	7	8	9	10
FLE*Months Since Reform	0.007	0.009	0.015	0.006	0.012
	[0.006]	[0.007]	[0.010]	[0.007]	[0.020]
Observations	1335	1335	1335	1335	1335
R-squared	0.93	0.94	0.94	0.95	0.96

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 "free lunch eligible" 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. 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 5.

Table 10: Heterogeneous Treatment Effects by Position on the Test Score Distribution, 4th Grade Math Scores

Panel A: Basic Set of Controls

Score percentile	*10th*	*25th*	*Median*	*75th*	*90th*
FLE*Year = 2000	0.604	0.851	0.884	0.854	1.198
	[0.94]	[0.90]	[0.77]	[0.70]	[0.73]
FLE*Year = 2003	3.047	2.817	2.417	2.271	2.707
	[0.74]**	[0.68]**	[0.63]**	[0.59]**	[0.69]**
FLE*Year = 2005	2.614	3.024	3.124	3.376	3.588
	[1.00]*	[0.94]**	[0.89]**	[0.87]**	[1.06]**
FLE*Months Since Reform	0.024	0.023	0.020	0.019	0.018
	[0.01]**	[0.01]**	[0.01]*	[0.01]*	[0.01]+

Panel B: Full Set of Controls

Score percentile	*10th*	*25th*	*Median*	*75th*	*90th*
FLE*Year = 2000	0.859	0.952	0.841	0.684	0.752
	[0.94]	[0.93]	[0.79]	[0.68]	[0.74]
FLE*Year = 2003	3.932	3.434	2.729	2.207	2.329
	[0.92]**	[0.86]**	[0.79]**	[0.73]**	[0.79]**
FLE*Year = 2005	3.68	3.823	3.59	3.208	2.902
	[1.14]**	[1.00]**	[0.83]**	[0.84]**	[1.01]**
FLE*Months Since Reform	0.032	0.027	0.020	0.013	0.011
	[0.01]**	[0.01]**	[0.01]*	[0.01]+	[0.01]

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

Note: For each panel, the first part corresponds to regressions in Table 5, and the second part corresponds to regressions in Table 6. The "basic" and "full" specifications correspond to the specifications in Column 3 and Column 5 of Table 5 (Table 6) respectively. See notes in Table 5.

Table 11: Heterogeneous Treatment Effects by Sex and Race, 4th Grade Math Scores

Panel A: Sex

Controls	*Basic*		*Full*	
Sex	Girls	Boys	Girls	Boys
FLE*Year = 2000	0.686	1.042	0.657	0.628
	[0.833]	[0.877]	[0.997]	[1.002]
FLE*Year = 2003	2.247	2.709	2.586	2.342
	[0.817]**	[0.612]**	[1.200]*	[0.821]**
FLE*Year = 2005	2.935	3.218	3.428	2.64
	[1.080]**	[0.932]**	[1.345]*	[1.074]*
FLE*Months Since Reform	0.020	0.020	0.022	0.014
	[0.009]*	[0.008]*	[0.010]*	[0.010]

Panel B: Race

Controls	*Basic*		*Full*	
Race	White	Nonwhite	White	Nonwhite
FLE*Year = 2000	1.037	-0.427	0.016	1.729
	[0.631]	[1.454]	[0.781]	[1.281]
FLE*Year = 2003	2.418	1.928	1.356	4.574
	[0.526]**	[1.185]$^{+}$	[0.691]$^{+}$	[1.326]**
FLE*Year = 2005	3.55	1.254	1.863	4.18
	[0.848]**	[1.539]	[1.004]$^{+}$	[1.493]**
FLE*Months Since Reform	0.026	0.006	0.018	0.02
	[0.007]**	[0.012]	[0.009]$^{+}$	[0.016]

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

Note: The same as in Table 10.

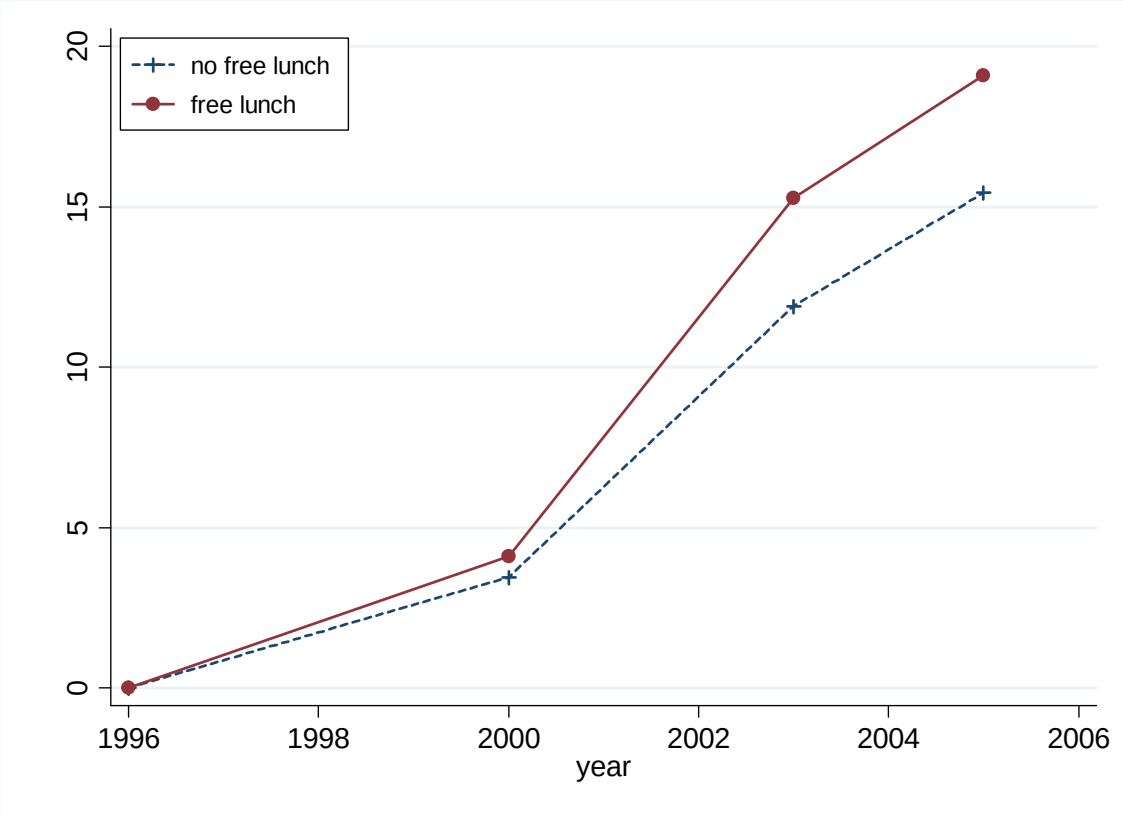
Table 12: Welfare Reform and 8th Grade Math Scores

	1	2	3	4
Controls	Basic	Full	Basic	Full
FLE*Year = 2000	-0.696	0.184		
	[0.915]	[0.879]		
FLE*Year = 2003	0.189	1.487		
	[0.820]	[0.986]		
FLE*Year = 2005	2.039	3.282		
	[1.044]+	[1.374]*		
FLE*Months Since Reform			0.003	0.007
			[0.009]	[0.012]
Months Since Reform			0.031	0.113
			[0.088]	[0.007]**
Year = 2000	0.245		-1.551	
	[1.348]		[4.095]	
Year = 2003	3.28		0.636	
	[1.708]+		[7.403]	
Year = 2005	1.84		-0.225	
	[2.220]		[9.507]	
White	19.757	19.641	19.932	19.795
	[0.758]**	[0.794]**	[0.710]**	[0.750]**
Male	0.642	0.626	0.632	0.597
	[0.228]**	[0.243]*	[0.224]**	[0.233]*
FLE	-17.054	-13.532	-17.611	-19.543
	[0.610]**	[8.671]	[0.535]**	[8.288]*
Accountability	2.488		1.138	
	[0.785]**		[0.781]	
Years of Accountability	0.544		0.453	
	[0.246]*		[0.248]+	
FLE*Accountability	-2.178	-1.99	-1.072	-0.928
	[0.943]*	[0.844]*	[0.798]	[0.716]
FLE*YrsAccount	-0.112	-0.05	-0.091	0.068
	[0.124]	[0.118]	[0.119]	[0.131]
Constant	231.389	168.868	209.129	152.499
	[39.281]**	[4.294]**	[30.673]**	[2.533]**
Observations	998	998	1255	1255
R-squared	0.94	0.95	0.93	0.95

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

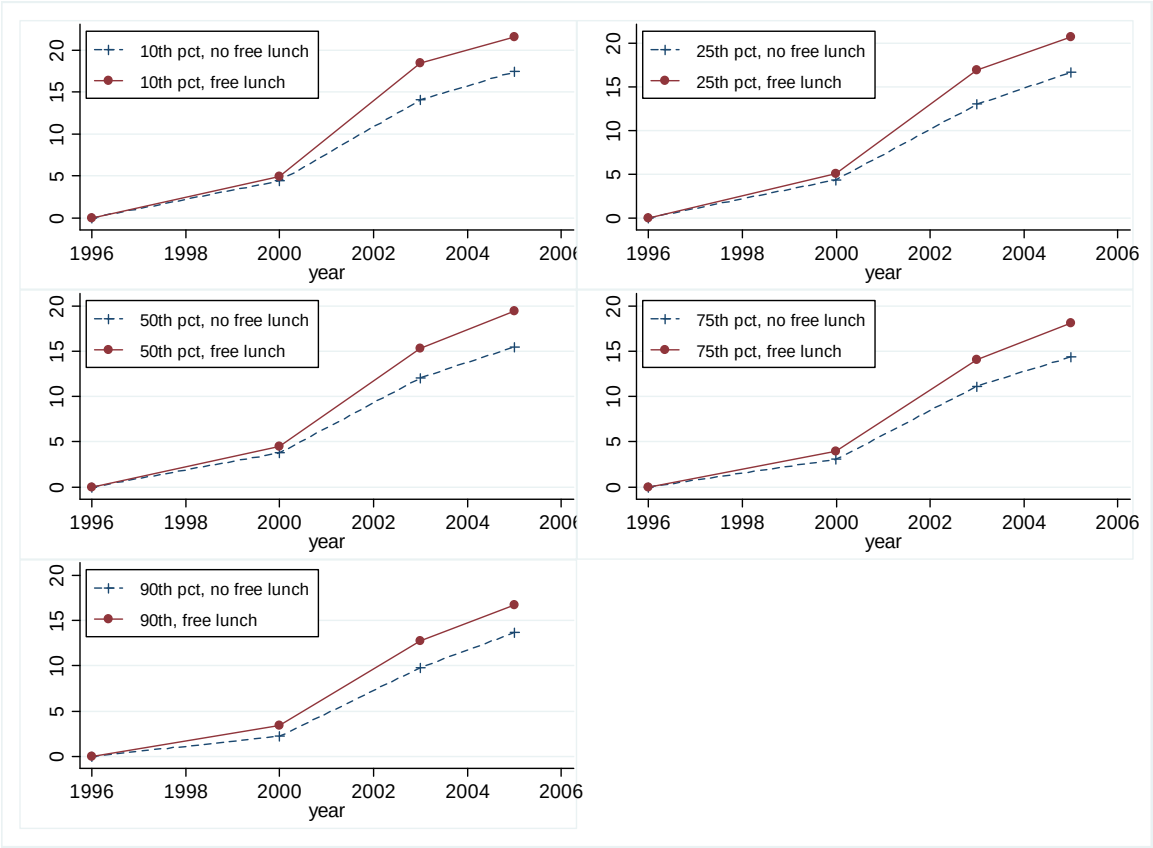
Note: The first two columns have the same specifications as those in Columns 3 and 5 of Table 5; the last two columns have the same specifications as those in Columns 3 and 5 of Table 6. See notes in Table 5.

Figure 1: 4th Grade Math Score Gains Relative to 1996 Values by Free-Lunch Eligibility



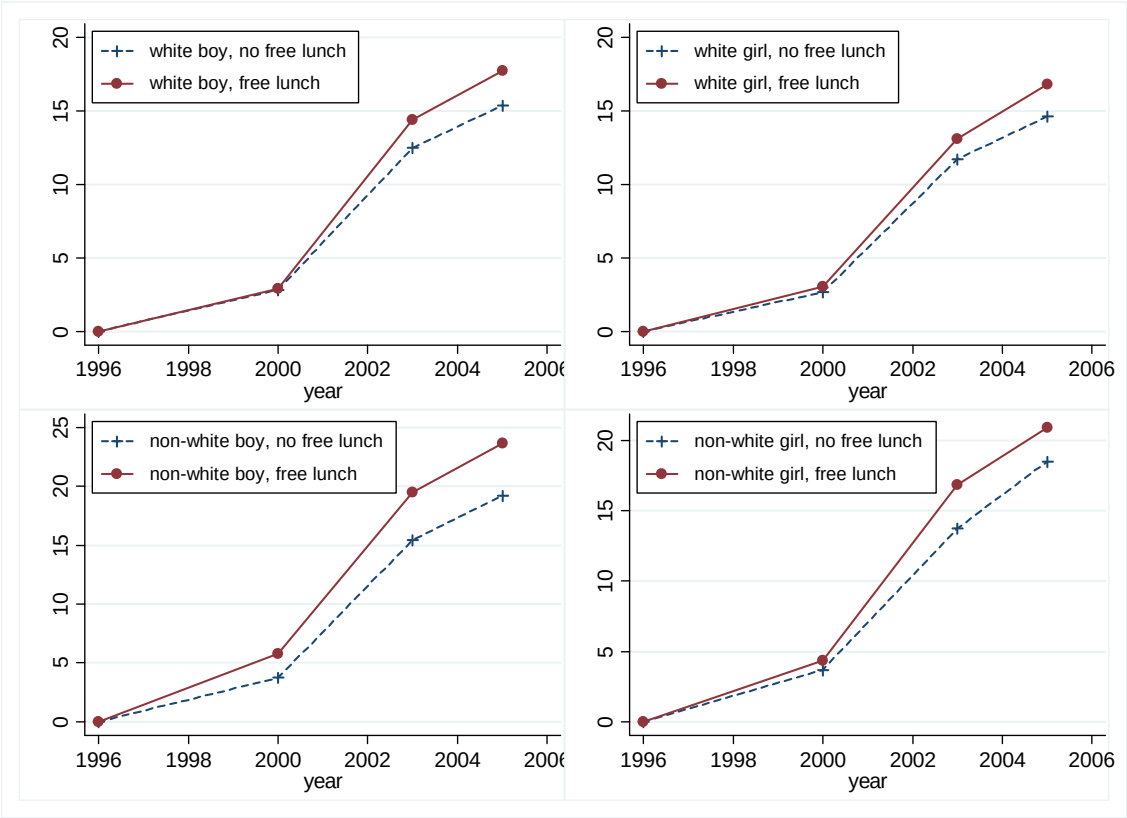
Sources: Authors' calculation from NAEP. The values for 2000, 2003, and 2005 are test score gains relative to the 1996 value within each free-lunch eligibility category. Test scores are measured on a 500-point scale.

Figure 2: 4th Grade Math Score Gains Relative to 1996 Values by Free-Lunch Eligibility and Position on Test Score Distribution



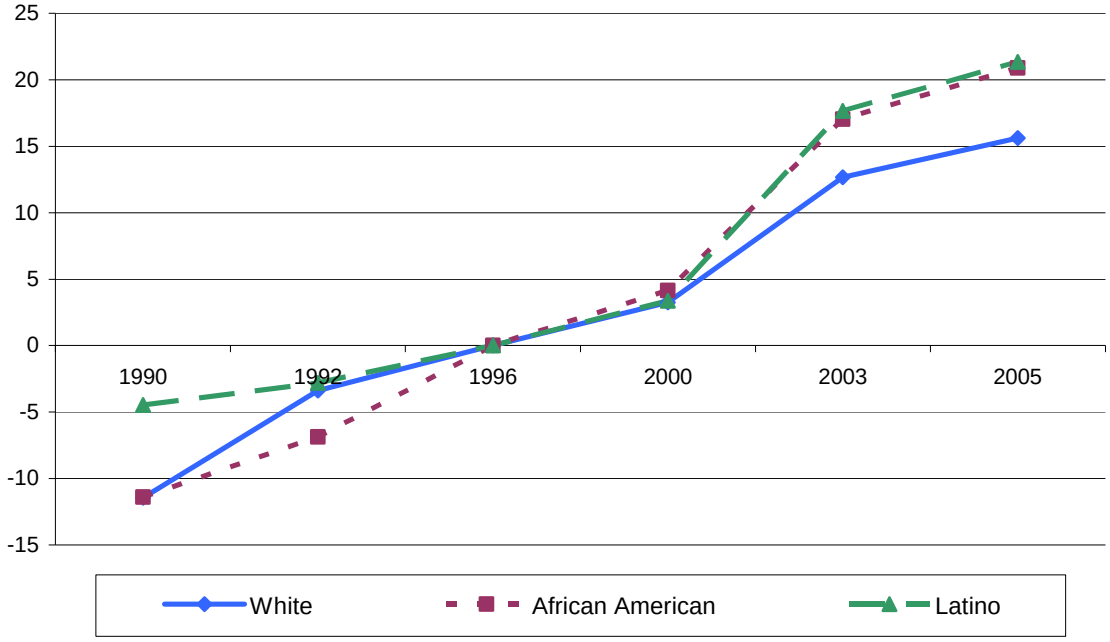
Sources: Authors' calculation from NAEP. The values for 2000, 2003, and 2005 are test score gains relative to the 1996 value within each free-lunch eligibility category. Test scores are measured on a 500-point scale.

Figure 3: 4th Grade Math Score Gains Relative to 1996 Values by Free-Lunch Eligibility, Race and Gender



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

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



Sources: 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

The size and direction of the bias in our estimates of the treatment effect (Table 5 and Table 6) will depend on the performance distributions of students in the three income groups: FLE, FLI, and NA. To illustrate, consider a simple difference-in-differences estimate with two time periods and three income reporting groups. Let P^0_{FLE}, P^0_{FLI}, 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: FLE*, and FLI*. If schools had full information about students' family incomes, the true test scores of the FLE and FLI groups would be $P^0_{FLE,FLE*}$ and $P^0_{FLI,FLI*}$. In the period following welfare reform, there is full information regarding income, and we observe $P^1_{FLE,FLE*}$ and $P^1_{FLI,FLI*}$. According to our model (ignoring the independent stochastic error terms),

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

$$P^1_{FLI,FLI^*} = P^0_{FLI,FLI^*} + t \text{,}$$

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

$$x_{obs} = (P^1_{FLE,FLE*} - P^1_{FLI,FLI*}) - (P^0_{FLE} - P^0_{FLI})$$
$$= x + [(P^0_{FLE,FLE*} - P^0_{FLE}) - (P^0_{FLI,FLI*} - P^0_{FLI})]$$

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

If assignment to NA in the initial period is random (uncorrelated with performance) within each of the two income groups, then $P^0_{FLE*}=P^0_{FLE}$ and $P^0_{FLI*}=P^0_{FLI}$, 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 FLE* group and the FLE group was more positive than the ability difference between the FLI* group and the FLI 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 FLE* performed better than FLE, and FLI* worse

than FLI. Although we never observe FLE* and FLI* 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 FLE or FLI. 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 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 FLE groups to produce the average pre-reform performance of the low-income group ($P^0_{FLE,FLE*,FLI*}$).[1] The counter-factual estimate of the treatment effect is:

$$x_{cf} = (P^1_{FLE,FLE*} - P^1_{FLI,FLI*}) - (P^0_{FLE,FLE*,FLI*} - P^0_{FLI})$$
$$= x + [(P^0_{FLE,FLE*} - P^0_{FLE,FLE*,FLI*}) - (P^0_{FLI,FLI*} - P^0_{FLI})]$$

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

The bias ($x_{cf} - x$) depends on the ability of FLI* students relative to the other groups, and the size of the FLI* group as a fraction of the *NA* group (f_{FLI}). Using the observed 1996 test scores and population shares, it is possible to determine the range of f_{FLI} for which the bias will always be negative under various assumptions about FLI* ability, and estimate an upper-bound on the magnitude of the positive bias. If FLI* baseline performance is at least as good as FLI performance ($P^0_{FLI*} \geq P^0_{FLI}$), the bias will be negative for all possible values of f_{FLI}. 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 ($P^0_{FLI*} \geq P^0_{FLE*}$), we can compute a maximum value for the bias as a function of f_{FLI}. For whites, the bias is negative for the full range from zero to one. For non-whites, the bias is negative for values of f_{FLI} greater than 14%. For

[1] 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 FLE average for all years, although, naturally, they constitute a shrinking share of the population of the group.

lower values of f_{FLI}, the bias ranges from zero to a maximum value of 0.0047.[2] This tells us that under the weak assumption that $P^0_{FLI*} \geq P^0_{FLE*}$, x_{cf} will be a lower-bound estimate of the treatment effect for whites, 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 $P^0_{FLI*} = P^0_{FLE*}$. In the more likely scenario that $P^0_{FLI*} > P^0_{FLE*}$, the bias will be more negative. The counterfactual estimates shown in Table 9 are smaller than the main estimates presented in Table 5, but are also substantially larger than 0.0047. Hence, even our *lower-bound* point estimates are positive, although not always statistically significant.

[2] Somewhat less informative, considering our unit of observation, is the value for the total population, which is negative for all values of f_{FLI}.

Appendix Table 1: Heterogeneous Treatment Effects by Position on the Test Score Distribution, Matched by Mean 1996 Scores

Panel A: Basic Set of Controls				
FLE Score Percentile	*25th*	*Median*	*75th*	*90th*
FLI Score Percentile	*10th*	*25th*	*Median*	*75th*
FLE*Year = 2000	0.988	0.455	0.2	0.216
	[0.911]	[0.833]	[0.677]	[0.649]
FLE*Year = 2003	2.02	1.514	1.369	1.142
	[0.677]**	[0.636]*	[0.516]*	[0.560]*
FLE*Year = 2005	2.579	2.274	2.39	2.465
	[0.875]**	[0.841]*	[0.750]**	[0.859]**
FLE*Months Since Reform	0.02	0.015	0.015	0.013
	[0.008]*	[0.008]*	[0.007]*	[0.008]

Panel B: Full Set of Controls				
FLE Score Percentile	*25th*	*Median*	*75th*	*90th*
FLI Score Percentile	*10th*	*25th*	*Median*	*75th*
FLE*Year = 2000	1.086	0.637	0.235	0.177
	[0.960]	[0.863]	[0.689]	[0.681]
FLE*Year = 2003	2.408	1.93	1.56	1.102
	[0.938]*	[0.833]*	[0.727]*	[0.751]
FLE*Year = 2005	2.932	2.736	2.539	2.311
	[1.070]**	[0.902]**	[0.753]**	[0.908]*
FLE*Months Since Reform	0.02	0.013	0.011	0.006
	[0.009]*	[0.008]+	[0.007]	[0.007]

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

Note: For each panel, the first part corresponds to regressions in Table 5, and the second part corresponds to regressions in Table 6. The "basic" and "full" specifications correspond to the specifications in Column 3 and Column 5 of Table 5 (Table 6) respectively. See notes in Table 5.