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

Amalia R. Miller
Lei Zhang

Abstract

During the 1990s, U.S. welfare policy underwent dramatic reforms aimed at promoting employment and reducing dependence. Although the immediate effects on adult labor supply and family income have been studied extensively, this paper is the first to evaluate the long-run effects on children's well-being. Using a decade of national math achievement data and controlling for contemporaneous changes in education policy and environment, we associate welfare reform with relative test score improvements for low-income students. Greater gains occur in states with larger initial welfare caseloads and larger caseload reductions.© 2009 by the Association for Public Policy Analysis and Management.

INTRODUCTION

During the early 1990s, the U.S. state welfare systems underwent the most dramatic reforms of the past half century. These reforms culminated in fundamental national reform in August 1996, when the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) replaced Aid to Families with Dependent Children (AFDC) with Temporary Assistance for Needy Families (TANF). The reforms aimed at promoting employment among recipients and reducing dependence on public assistance through time limits, job subsidies, work requirements, and increased funding for child care. While studies of welfare reform's short-term impacts on income and employment abound, little is known about its potential long-term, intergenerational effects. This paper evaluates the impact of welfare reform on the academic performance of children in low-income families. The educational outcomes of children are meaningful indicators of their long-term economic and social well-being, and hence constitute a key criterion by which scholars can assess the long-term effects of welfare reform.

This study is the first to measure the effects of national and statewide welfare reforms on the academic performance of low-income children. We use individual math achievement scores from the National Assessment of Educational Progress (NAEP) covering the decade following PRWORA to estimate the net effect of welfare reform. Empirically isolating the impact of welfare policy is inherently challenging. The reforms under study are not an ideal natural experiment, as they were enacted over a period of several years that also involved numerous economic and policy changes. This paper pursues several complementary strategies to control for confounding factors and measure the impact of welfare reform on low-income students.

Journal of Policy Analysis and Management, Vol. 28, No. 4, 577–599 (2009)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience
(www.interscience.wiley.com)
DOI: 10.1002/pam.20456

We start by measuring the impact of national reform through PRWORA, focusing on states that had not previously adopted significant welfare reforms. Estimation is conducted in a difference-in-differences framework, with students from more affluent families serving as the control group. We control for time-varying educational inputs, such as school spending and parental education, allowing for differential effects on low-income students. We also include a rich set of state and time fixed effects to absorb unobservable factors that affect student performance. We find no evidence that national welfare reform harmed the academic performance of students in low-income families. On the contrary, math scores of low-income 4th graders grew more rapidly in the post-reform era than scores of comparable higher-income 4th graders. The estimated effects are statistically and economically significant starting in 2000, and are larger in states with higher initial caseloads or larger caseload reductions. We confirm the robustness of the estimates after controlling for contemporaneous policy changes that may have affected low-income students differentially, such as the adoption of school accountability programs and the expansion of the Earned Income Tax Credit (EITC). We also associate national welfare reform with significant relative test score growth for low-income students between 4th and 8th grades. These results indicate persistence of the gains accrued at younger ages. Finally, we exploit cross-state variation in the timing of state reforms to link the reforms to test score gains in all states.

The existing welfare reform evaluation literature has focused primarily on outcomes for adults and households. Blank (2002), Moffitt (2003), and Grogger and Karoly (2005) provide excellent reviews of this literature. In the years following state and national welfare reforms, there was a sharp decline in welfare caseload and a large increase in the labor supply of single mothers with children under 18. There is some controversy regarding the effects of welfare reform on family income. While some find evidence of rising incomes for female-headed households, others document income declines for low-education families and a rise in "deep poverty." Meanwhile, studies show consumption did not decline for vulnerable women as a result of welfare and tax reforms. Evidence on family structure is not conclusive and the effects have not been demonstrated with national data.

The most compelling previous studies of welfare reform's impacts on children employed an experimental framework with random assignment into treatment (welfare reform) and control (traditional welfare) groups, using data from five welfare demonstrations. These studies find evidence of improvements in self- and teacher-reported academic achievement and test scores for younger children in the treatment group 2 to 5 years after exposure to welfare reform. For older children and adolescents, however, they find evidence of reduced academic performance, measured by self-reported dropout, expulsion and suspension rates, and maternal assessments of overall achievement (Duncan & Chase-Lansdale, 2001; Gennetian et al., 2004; Morris, Duncan, & Clark-Kauffman, 2005). Although proper randomization eliminates self-selection bias, these experiments were limited geographically, tended to employ weaker work requirements than those in statewide and national programs, and could not capture entry effects due to welfare participation. Therefore, their external validity may be compromised. Using observational data on a nationally representative sample of children to study actual statewide and national welfare policy changes, with a study design that incorporates potential entry effects, this paper confirms the positive effects of welfare reform found in the experimental literature. The results in this paper provide substantial support for the experimental finding of short-term educational benefits from welfare reform for young children. Additionally, we extend the time horizon beyond previous studies and show longer-term educational gains more than a decade following welfare reform.

EMPIRICAL FRAMEWORK

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

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

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. This framework includes inputs from formal education as well as home production and allows for current performance to depend on current and past inputs. In particular, family inputs are assumed to contribute to human capital once a child is born, while school inputs accumulate from the age of school entry.

The primary channel through which welfare reform can affect the school performance of low-income students is by altering family inputs and student effort. On the one hand, increased maternal employment may hinder child cognitive development by reducing time available for home production, such as supervising and disciplining children, reading to them, and assisting with homework. When maternal time with children decreases, there may be a substitution toward alternative caregivers, such as family members, formal child care providers, and structured school-based programs, depending on the age of the child. On the other hand, working mothers may feel more secure and confident or they may learn child-rearing strategies from coworkers, resulting in greater productivity at home. Working mothers may also provide their children with improved stability and daily routine, serve as positive role models for their children, and instill a desire for financial independence and greater academic achievement. Increased family income may also boost children's school performance through improved nutrition and reading materials at home, although the earned income gains were largely offset by reduced welfare benefits.[1] Welfare reform may also directly affect the incentives to invest in human capital, by reducing the availability and attractiveness of welfare as a long-term substitute for paid employment. These different channels operate in different directions, and the net effect of welfare reform on school performance is inherently an empirical question.

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

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

where s indexes state, t indexes time, and i indexes the individual student. We assign students to the low-income treatment group based on their eligibility for federal (USDA) subsidized lunches. $FLE_i = 1$ if i is eligible for free school lunches and $\tau_t = 1$ if t is a post-reform year. X_{ist} includes proxies for both school inputs, such as staffing and expenditures, as well as family economic inputs, such as adult educational attainment, income, and employment. To account for the cumulative nature of the production function, whereby current test scores are affected by past investments in human capital, the home controls are measured as averages

[1] Prior studies of maternal employment and child human capital produce mixed evidence. Negative effects are concentrated among children in more affluent, two-parent families, with more educated mothers (Baum, 2003; Blau & Grossberg, 1992; Gregg et al., 2005: Ruhm, 2004). For financial resources, Blau (1999) finds only trivial direct effects of family income on child development. Waldfogel (2007) reports no positive association between U.S. welfare reform and low-income family spending on education or children's clothing.

between the birth year and the test year, and the school inputs are averaged from age six to the test year.

Our parameters of primary interest are the year-eligibility interaction terms, $\beta_{FLE,t}$. These difference-in-differences (DD) estimates of the net treatment effect capture the difference in test score growth between the FLE group and the higher-income FLI (free lunch ineligible) control group in the period following welfare reform. The assumption is that test score changes for higher-income students provide a counterfactual for what would have happened to test scores of low-income students absent welfare reform.

The income cutoff for FLE is 130 percent of the federal poverty level. Hence, the treatment group is a superset of the welfare-eligible that likely contains all formerly and currently eligible children, as well as children from somewhat wealthier families that are susceptible to welfare receipt.[2] If the low-income proxy for the welfare eligible is overly inclusive or if there exist positive cross-income peer spillovers, the $\beta_{FLE,t}$ estimate will be attenuated. We calculate from the Survey of Income and Program Participation (SIPP) that 25.4 percent of FLE children received welfare payments in 1996, while virtually none of the control group received such payments. Another advantage of using free lunch eligibility rather than welfare receipt as the low-income proxy is the relatively stable composition of the treatment and control groups over time. We calculate from the SIPP that 23.1 percent of all children below 18 years of age were eligible for free school lunches in 1996. By 2007, the share had dropped only slightly to 21 percent.

We estimate cumulative treatment effects over different time periods following welfare reform. There are two main reasons we expect these estimates to increase with time. The first relates to the *duration* of exposure to welfare reform leading up to each test date. Test scores measure the current stock of knowledge accumulated through past investment, less depreciation. If the impact of welfare reform is to increase investment by a fixed amount per year, for example, the total impact on scores will increase linearly with years of exposure.[3] The second reason to anticipate a growing treatment effect is that *age at first exposure* to welfare reform may influence the human capital investment response. For example, maternal employment may have different effects on preschool-aged children and adolescents. Furthermore, the positive educational effects found in the experimental literature on welfare reform are primarily for younger children. The 4th-grade students in our sample were initially exposed to welfare reform at around age 10 (for those tested in 1996), age 6 (tested in 2000), age 3 (tested in 2003), infancy (tested in 2005), and at birth (tested in 2007). Since duration of exposure to welfare also increases across cohorts, it is impossible to disentangle these two hypotheses using 4th-grade tests alone. The same problem applies to 8th-grade students, whose age at initial exposure ranges from 14 (for those tested in 1996) to 4 (tested in 2007). We draw on scores for both grades below to explore this issue.

After estimating the basic DD treatment effects of welfare reform, we extend the analysis to address potential sources of bias. We account for other time-varying factors that may cause differential changes in test scores for low- and high-income groups. These variables may be socioeconomic or political. Given the cumulative nature of the education production process, these confounding factors include changes that occurred before or after PRWORA. We identify and consider two important policy changes: (1) state-level school accountability reforms starting in the early 1990s and

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

[3] This is similar to the approach in Bleakley (2007) that allows the impact of malaria eradication efforts to vary with years of childhood exposure. Another duration-related theory is that there may be a lag between welfare reform and the full response in family inputs, for example, because time limits are not immediately binding.

the passage of the No Child Left Behind Act (NCLB) in 2002, and (2) the expansions of Earned Income Tax Credit (EITC) during the late 1980s and the early 1990s. The estimated treatment effects are robust to controlling for these policy changes.

DATA

School Performance

We measure school performance with test scores from the Department of Education's National Assessment of Educational Progress (NAEP) program.[4] The 500-point tests are given to a representative sample of about 3,000 students in each state between late January and early March of the test year. The NAEP provides an independent measure of knowledge and aptitude of U.S. students in mathematics and has been widely used for cross-state and time-series comparisons of student performance.[5]

We use confidential Restricted Use micro-level data from NAEP mathematics assessments in 1996, 2000, 2003, 2005, and 2007. These data sets are available by request from the Department of Education. The crucial benefit of the confidential data set is that it distinguishes between free lunch eligibility and reduced-price lunch eligibility, allowing us to define a more appropriate treatment group.[6] The first year of our study is 1996, the year in which the NAEP began collecting student-level information on low-income status. A small and declining fraction of NAEP test-takers is not assigned to an income category due to changes in income reporting. We omit these students from our empirical analysis. In a separate exercise, we test the potential bias from changes in the composition of the FLE and FLI groups caused by changes in income reporting rates. We produce lower-bound estimates that confirm the main results (Miller & Zhang, 2008b).

This paper focuses on math test scores due in part to the importance of math scores for predicting economic outcomes and in part to the limited range of years available for other NAEP subject tests. The NAEP tests in reading and science are only available by free lunch eligibility status starting in 1998 and 2000, respectively. This makes it impossible to conduct our full empirical analysis in those subject areas. Using the limited time periods, and the same sets of fixed effects and control variables used in the main analysis, 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, but not always statistically significant. Comparing science scores in the two available years (2000 and 2005), we also find relative gains for the FLE group, generally statistically significant. These relative gains in reading and science suggest that the observed gains in math were not achieved at the expense of other subject areas.

Summary statistics of test scores by income and year are reported in Table 1 of the online Appendix.[7] In Figure 1 of the paper, 4th-grade test score changes are

[4] Test scores are shown to have a significant effect on wages even after controlling for education attainment (Hanushek & Zhang, 2006; Murnane, Willet, & Levy, 1995).

[5] Compared to other Department of Education survey datasets, NAEP has the advantage that it provides comparable test scores for a repeated cross section of students in elementary and middle schools over the 1990s and 2000s, the period in which welfare reform occurred. Among recent longitudinal survey data, the Education Longitudinal Study of 2002 (ELS) follows 10th graders in 2002, while the Early Childhood Longitudinal Study (ECLS) follows kindergarteners in 1998 and infants in 2001. Neither provides enough academic achievement information to allow meaningful comparison over the period of welfare reform. Additionally, they have much smaller samples than the NAEP. The major drawback of NAEP is the lack of individual background information such as parental education and income.

[6] It is possible to conduct a similar analysis using only publicly available NAEP data with an aggregate unit of observation and a low-income treatment group defined by subsidized lunch eligible, rather than free lunch eligible. The estimated effects on the broader treatment group are smaller than, but qualitatively similar to, the results in this paper. See Miller and Zhang (2008b) for complete results using publicly available data.

Table 1. National welfare reform and 4th-grade math scores.

Model	No School Accountability Reform	Basic	Full	Share Near Proficient	Share Near Proficient (IV)
State*Year FE	N	N	Y	Y	Y
FLE*Controls	N	N	Y	Y	Y
	1	2	3	4	5
FLE*Year2000	1.786*	4.796**	4.368**	4.395**	3.177+
	(0.681)	(0.955)	(0.960)	(1.058)	(1.615)
FLE*Year2003	2.500**	5.299**	5.123**	5.355**	4.159*
	(0.561)	(0.922)	(1.097)	(1.146)	(1.655)
FLE*Year2005	3.163**	9.092**	9.089**	9.87**	8.645**
	(0.541)	(0.765)	(0.924)	(1.019)	(1.519)
FLE*Year2007	2.776**	8.804**	9.145**	9.877**	8.503**
	(0.699)	(0.914)	(1.018)	(1.112)	(1.536)
Share near proficient				0.088*	0.102
				(0.041)	(0.116)
Observations	525,920	525,920	525,920	525,920	295,030

Notes: Test scores are from 1996–2007 NAEP tests in states without statewide welfare reforms in place prior to 1995. The unit of observation is an individual student in a given year. The low-income category is FLE = free lunch eligible (income below 130 percent of the poverty level). The number of observations is rounded to tens.

All regressions include controls for sex, race, income category, and time-varying controls for the state economy and school inputs. The controls are interacted with FLE in columns 3–5. School accountability measures are in columns 2–5. Columns 4 and 5 include controls for NCLB intensity, measured as share of students in each income group whose 2003 state math test score fell into either proficient or basic categories (share near proficient). In column 5, the share near proficient is treated as endogendous. All regressions include state and year fixed effects. Columns 3–5 include the full set of fixed effects for State*Year interactions.

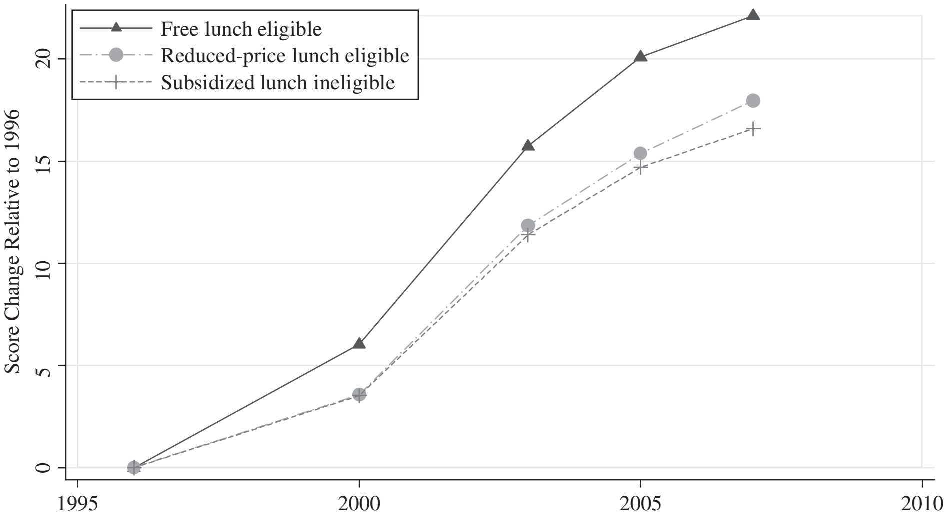
Some coefficients suppressed for readability. Full results for the main estimates (columns 2 and 3) are provided in the first two columns of Table 4 in the online Appendix. Results from additional robustness tests are in Table 5 of the online Appendix (go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787).

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

depicted as deviations from 1996 baseline scores. Over the period between 1996 and 2007, there has been an across-the-board improvement in math test scores. The pattern of test score gains, however, differs substantially between the FLE and FLI groups, with the latter comprising SLI (subsidized lunch ineligible, income above 185 percent of poverty) and RLE (reduced-price lunch eligible only, income between 130 and 185 percent of poverty) students. The gains for the FLE group are substantially larger than those for both the SLI and the RLE groups for the entire period, and the difference is largest for 2007. The RLE test score trend, by contrast, is similar to the SLI trend. Given the timing of national welfare reform and the differential welfare rates between the FLE and FLI groups described in the empirical framework, this relative increase in test scores for FLE children may be attributable to the impact of welfare reform on education production.[8] The regression analysis that follows aims to eliminate potentially confounding factors.

[7] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

[8] According to SIPP data, the welfare receipt rate dropped from 25.4 percent to 9 percent between 1996 and 2007 among FLE children, whereas virtually no FLI children received welfare in either year.



Source: NAEP Restricted Use State Mathematics Assessments: 1996, 2000, 2003, 2005, and 2007. Figure plots test score gains relative to the 1996 value within each income category. Test scores are measured on a 500-point scale. Sample consists of all states.

Figure 1. 4th-Grade Math Score Deviations from 1996 by Free and Reduced-Price Lunch Eligibility.

Welfare Reform

Prior to PRWORA, several states adopted statewide welfare reform under AFDC waivers at various dates starting in 1992. The main analysis in the next section uses a DD approach to study the impact of national welfare reform on children in all states without statewide waivers in place by 1995. In the following section, we exploit both cross-sectional and time-series variation in state welfare policies, which varied across states at any point in time and also evolved within states over time.

Data on timing of welfare policy changes are from Crouse (1999), U.S. DHHS (1997), and the Urban Institute Welfare Rules Database, and are summarized in Table 2 of the online Appendix.[9] The table reports the adoption timing of statewide welfare waivers, time limits for welfare receipt, sanctions for violations of work requirements, and school requirements for dependent children. Eleven states implemented statewide waivers before 1995. By 2005, all but 5 states had adopted a time limit policy: 14 initially imposed an intermittent time limit, which includes a periodic time limit, benefit reduction, or benefit waiting period; 27 imposed a lifetime time limit; and 5 states used both. Over time, 12 states increased the severity of their time limit policy, and 3 states reduced it. Nearly all states with a lifetime time limit policy used 60 months as the limit for the entire sample period. We define four types of sanctions based on the reduction in cash benefits the first time work requirements are violated (initial sanction) and on the most severe sanction for violation. Over time, 22 states switched to more stringent sanctions, and no state reverted to less stringent ones. By 2007, 32 states had adopted a school requirement

[9] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

Table 2. Heterogeneous treatment effects by sex and race.

Model	Basic	Full	Basic	Full
Sex*Year, Race*Year FE	N	N	Y	Y
	1	2	3	4
FLE*Year2000	2.961**	3.242**	3.232**	3.55**
	(0.963)	(0.988)	(1.006)	(1.052)
FLE*Year2003	1.378	2.119+	2.503*	3.244*
	(1.048)	(1.219)	(1.082)	(1.261)
FLE*Year2005	4.812**	6.037**	6.115**	7.31**
	(0.920)	(1.044)	(0.908)	(1.018)
FLE*Year2007	4.477**	5.741**	5.64**	6.854**
	(0.855)	(1.034)	(0.899)	(1.057)
Male*FLE*Year2000	0.915	0.865	0.502	0.383
	(0.603)	(0.599)	(0.672)	(0.670)
Male*FLE*Year2003	1.376**	1.300**	1.069*	0.953*
	(0.389)	(0.380)	(0.464)	(0.457)
Male*FLE*Year2005	1.106*	1.041*	1.344*	1.223*
	(0.495)	(0.484)	(0.537)	(0.519)
Male*FLE*Year2007	0.925*	0.864+	1.327**	1.223**
	(0.448)	(0.429)	(0.451)	(0.431)
NonWhite*FLE*Year2000	2.572*	1.725+	1.902*	1.260
	(0.969)	(0.909)	(0.886)	(0.935)
NonWhite*FLE*Year2003	6.184**	5.089**	1.777*	1.018
	(0.747)	(0.827)	(0.826)	(0.958)
NonWhite*FLE*Year2005	6.939**	5.409**	1.361+	0.292
	(0.637)	(0.643)	(0.713)	(0.725)
NonWhite*FLE*Year2007	7.151**	6.01**	1.777*	1.095
	(0.695)	(0.702)	(0.787)	(0.787)
Observations	525,920	525,920	525,920	525,920

Notes: Columns 1 and 2 estimate common time trends for sex and race groups, while columns 3 and 4 allow the year effects to vary by sex and race. The basic set of controls is used in columns 1 and 3 (corresponding to Table 1, column 2); the full set (Table 1, column 3) is used in columns 2 and 4. See Table 1 for further notes.

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

policy, which may include requirements such as children attending school, achieving a minimal grade point average, and other forms of parental involvement in their children's education.

Time-Varying State Characteristics

We use state characteristics to account for the major categories of time-varying educational inputs. Pupil–teacher ratio and expenditure per student are proxies for measurable school inputs; average income, unemployment rate, percentage of population 25 years of age or older with a high-school diploma or higher, and percentage with a bachelor's degree or higher reflect forces that may affect family inputs. School input variables are averages over the years a typical 4th grader is in school: 1992–1995, 1996–1999, 1999–2002, 2001–2004, and 2003–2006 for test years 1996, 2000, 2003, 2005, and 2007, respectively; family input variables are averages between the birth year of a typical 4th grader and the test year: 1987–1995, 1991–1999, 1994–2002, 1996–2004, and 1998–2006. School districts with high concentrations of low-income students may differ from other districts in both levels of

Table 3. Heterogeneous treatment effects by position on the test score distribution.

Percentile	10th	25th	Median	75th	90th
FLE*Year2000	8.21**	8.11**	8.31**	7.11**	6.11**
	(0.44)	(0.36)	(0.34)	(0.37)	(0.45)
FLE*Year2003	8.47**	7.67**	7.57**	6.27**	5.07**
	(0.35)	(0.30)	(0.30)	(0.29)	(0.36)
FLE*Year2005	14.79**	14.79**	14.79**	13.29**	12.09**
	(0.35)	(0.31)	(0.29)	(0.28)	(0.34)
FLE*Year2007	14.39**	14.79**	15.29**	13.69**	12.39**
	(0.36)	(0.31)	(0.30)	(0.28)	(0.37)
Observations	525,920	525,920	525,920	525,920	525,920

Notes: All regressions include the basic set of controls in Table 1, column 2. Estimates are obtained using the "changes-in-changes" (CC) estimator (Athey & Imbens, 2006). All standard errors are computed using bootstrapping with 1000 replications.

+ significant at 10%; * significant at 5%; ** significant at 1%.

and changes in educational inputs, partly resulting from nationwide school finance equalization since the early 1970s. To address this concern, we compute separate average state*year school spending and pupil–teacher ratios for FLE and FLI students, weighting district-level data by the number of students in each income group.

Table 3 in the online Appendix[10] reports summary statistics of these variables for each test year. During the sample period, there is an overall increase in school inputs reflected by increasing expenditures per student and decreasing pupil–teacher ratios. State income and education levels have also increased over the period. Using SIPP data, we confirm these gains for both FLE and FLI households. The income gains for FLE households were absolutely and proportionally smaller ($18 per month between 1996 and 2007, in 1996 dollars) than those for FLI households ($591 per month). The gains in educational attainment were similar between FLE and FLI at lower levels (the fraction with less than a high school diploma dropped by about 40 percent for each group), but greater for FLI at higher levels. The fraction with a college degree increased 20 percent for FLI, but only 11 percent for FLE. We are unable to control for these factors at the individual level using NAEP data. To the extent that FLI students experienced relative improvements in inputs, our estimates will understate the benefits to FLE students from welfare reform.

Introduction of State School Accountability Systems

States began introducing school accountability systems in the early 1990s, and their effect on academic performance is not likely captured by school inputs. The uneven introduction of accountability systems over time and the differences in timing of the two types of reforms allow us to separately identify the effects of accountability systems and welfare reform on school performance (see Table 2 in the online Appendix[11] for dates). Following Hanushek and Raymond (2005), we define an accountability system as a mechanism for publicly disseminating information on standardized test performance for each school, along with a way to aggregate and

[10] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

[11] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

interpret the school performance measure. States are classified as "consequential" states if they both report results and attach consequences to school performance, or "report card" states if they only provide a public report. States introduced their accountability systems between 1993 and 2003. By 2003, 31 states had consequential accountability.[12]

The passage of the federal No Child Left Behind Act (NCLB) in January of 2002 demanded stronger accountability of schools in all states. Between January and June of 2003, all states submitted their required plans for implementing an accountability system under NCLB to the Department of Education. These were approved by June 2003, after the 2003 test date. Therefore, we consider 2005 the first test year in which the 19 report card states had consequential accountability in place. We define a dummy variable equal to 1 if an accountability system was in place prior to an NAEP test year, and we create a variable for the years elapsed since a state implemented an accountability system. To allow for differential effects of accountability on FLE students, we also interact these two variables with the FLE indicator. Below, we explore the role of school accountability in detail and relax the assumption that NCLB was equivalent to a consequential state system.

EFFECTS OF NATIONAL WELFARE REFORM

This section establishes the main results by comparing test score growth for low-income students before and after national welfare reform to growth for higher-income students in the same period. The experimental literature found that welfare reform affected low-income students within 2 to 3 years after implementation. Hence, we exclude states with waivers in place before 1995.[13] The 1996 tests were conducted prior to PRWORA and serve as the baseline.

Main Estimates of the Average Effects of National Reform

Table 1 reports the effects of national welfare reform estimated from Equation (2). All specifications include state and year fixed effects as well as controls for student race, sex, and income group. All coefficients are reported with robust standard errors clustered at the state level, allowing for arbitrary error term heteroskedasticity and correlation within a state. All reported results are from unweighted regressions. Results are unchanged if we use NAEP sampling weights.

Column 1 of Table 1 shows treatment effect estimates for the FLE group in a model with only demographic controls, state and year fixed effects, and key time-varying inputs: state-year adult population shares with high school diplomas and college degrees, per capita income (in tens of thousands of 1983 dollars) and unemployment rates, public school expenditures per pupil (in thousands of 1983 dollars), and pupil–teacher ratios. The positive and significant interaction terms indicate test score gains for free lunch eligible students, relative to ineligible students, ranging from 1.8 to 3.2 test score points. The significantly positive estimate for year 2000, the first test after the 1996 baseline, suggests immediate short-term test score gains. The magnitude of the treatment effect estimates increases after 2000, although the later years may be more susceptible to external confounding influences (such as the No Child Left Behind Act). Other regression coefficients (unreported in the table) indicate that males outperformed females by 2 points, whites outperformed non-whites by 13 points, and FLE students lagged by 20 points. The positive and significant year

[12] Implementation dates of state school accountability reforms are from Hanushek and Raymond (2005), Fletcher and Raymond (2002), and Goertz and Duffy (2001).

[13] Results are robust to using 1994 or 1996 as the cutoff year. Analysis in this and the remaining sections excludes the District of Columbia. DC has heavy concentrations of low-income and minority students and has long stood out on the NAEP tests for its poor performance. Including DC increases the magnitude and significance of the treatment effect estimates.

Table 4. Value-added estimates for test score growth between 4th and 8th grades.

Model	Basic	Full	Full
Sample	All	All	Non-White
	1	2	
FLE*Born1991	1.082	1.807*	3.159**
	(0.721)	(0.863)	(1.105)
FLE*Born1994	1.523+	2.494*	4.879**
	(0.859)	(1.100)	(1.425)
Average 4th-grade score	0.468**	0.502**	0.255**
	(0.056)	(0.058)	(0.089)
Observations	792	792	396

Notes: Selected regression estimates from the value-added model (Equation 3). The omitted birth year cohort is 1987. Unit of observation is the State*Year*Race*Sex*FLE cell, aggregated from Restricted Use Data. Sample is limited to non-waiver states. Robust standard errors clustered at state level in parentheses.

+ significant at 10%; * significant at 5%; ** significant at 1%.

effects reflect the overall improvement in test scores during the period. A falsification test estimating the treatment effects on students eligible only for reduced-price lunch (RLE) shows that, in contrast to the FLE group, the RLE*year interactions are substantially smaller (ranging from 0.2 to 1.9) and mostly statistically insignificant.

In columns 2 and 3 of Table 1, we address the potential role of state school accountability reforms in the relative test score gains for low-income children. In column 2, we include an indicator for the presence of a consequential accountability program, a linear measure of years since the accountability program was initiated, and interactions of FLE with each of these.[14] This is our basic model. Accountability has a positive and significant effect on test scores, and years since accountability is also positive but insignificant (see Table 4 of the online Appendix[15] for complete results). The magnitude of the estimates is comparable to Hanushek and Raymond (2005). The FLE–accountability interaction is negative and significant. As a result, controlling for school accountability actually increases the size of the treatment effect estimates, which now range from 4.8 to 9.1 points. Column 3 reports results from our fully interactive model, with a complete set of fixed effects for state and year interactions to remove any state-specific time trends in test scores. The model also includes interactions between the time-varying state educational inputs and the FLE indicator, allowing them to have differential effects on low-income students. The FLE interactions with the adult population share that is college educated and the average pupil–teacher ratio are each negative and significant. The main variables of interest are again positive, statistically significant, and increasing between 2000 and 2005, and leveling off in 2007.

Welfare reform has a nontrivial effect on test scores, irrespective of model specification. Consider the estimates from our basic model reported in column 2. By 2000, within 4 years after the 1996 national reform, 4th-grade FLE students increased their NAEP math scores by 4.8 points more than FLI students. This additional growth corresponds to 0.16 standard deviation of the initial 1996 test score distribution (see Table 1 of the online Appendix[16]). This value is within the range of

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

[15] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

[16] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

magnitudes in the experimental studies, which report gains for the treatment group of 0.04 to 0.25 standard deviations within 3 to 5 years of randomization (Duncan & Chase-Lansdale, 2001). Between 1996 and 2007, FLE students increased their relative test scores by 9 points, corresponding to 75 percent greater test score growth than FLI students. This additional growth from a decade of welfare exposure represents 0.3 standard deviation of the initial 1996 test score distribution. A useful comparison for these effects is the 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 test scores by 0.17 standard deviation in kindergarten. 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, the estimated impact of welfare reform on low-income students is quantitatively, as well as qualitatively, significant. Our estimates associate national welfare reform with a 13 to 40 percent narrowing of the test score gap between FLE and FLI students by 2007.

A consistent pattern across specifications is that the treatment effect estimates increase substantially between 2000 and 2005 and then remain stable. As discussed in the empirical framework, this dynamic response may be the result of greater years of exposure leading to larger cumulative effects or the result of early childhood exposure having a larger impact than later exposure. Between 1996 and 2005, each successive cohort of 4th-grade NAEP participants had been exposed to welfare reform for an increasing number of years leading up to their test date. The successive cohorts had also been initially exposed to welfare reform at younger and younger ages. Cohorts tested in 2005 and later were exposed to welfare reform from infancy.

In order to disentangle the age and duration theories, we add information from the 8th-grade NAEP mathematics tests. These tests were conducted in the same years as the 4th-grade tests. We estimate a modified version of Equation (2) for 8th-grade scores, using control variables averaged over the appropriate years for these older children. Table 6 of the online Appendix[17] shows results from the basic and fully interactive models. The FLE*Year interactions are positive, significant, and growing from 2000 to 2007. The relative test score gains by 2007 amount to 9 points, or 0.27 standard deviation of the test score distribution, slightly lower than the normalized gains in 4th grade.

Note that the 8th graders tested in 2000 were initially exposed to welfare reform by age 10. The significant relative test score gains experienced by 8th graders in 2000, together with the comparable magnitudes between the 4th grade and 8th grade gains between 1996 and 2000, suggest that welfare reform can improve outcomes for children with initial exposure in elementary school. This result contrasts somewhat with the experimental findings, where the benefits of reforms only accrued to children age 5 and younger at the time of randomization (Gennetian et al., 2004; Morris, Duncan, & Clark-Kauffman, 2005). For later test years, the 8th graders tested in 2003 were exposed to welfare reform by age 7, and those tested in 2007 were initially exposed by age 2. This evidence supports a role for a cumulative effect that increases with duration of exposure, but does not rule out larger effects from earlier exposure. Comparing the treatment effects between 2005 and 2007 provides additional support for the importance of duration or age at initial exposure. For low-income 4th graders, the maximal gains are achieved by 2005, the first test year in which children were exposed to welfare reform from infancy. For 8th graders, the estimated gains continue to increase in magnitude between 2005 and 2007.

[17] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

Robustness Tests for the Average Effects of National Reform

We first attempt to disentangle the effects of NCLB and welfare reform further. It is worthwhile to note that significant test score gains are detected in 2000, well before the passage of NCLB. However, for scores in 2005 and 2007, NCLB is a potential confounding influence. The Act required all states to disaggregate test results by socioeconomic group, something not universally required under earlier state school accountability systems. Therefore, NCLB may have had a stronger differential effect on FLE and FLI students than state accountability.

To isolate the additional differential effect of NCLB, we would ideally like to control for the differential pressure of NCLB on FLE and FLI students or their schools. A direct measure being unavailable, we use as a proxy the share of 4th graders whose math scores in 2003 state tests fell into one of the achievement categories adjacent to the proficiency cutoff—proficient or basic—for each income category and each state.[18] The hypothesis is that NCLB pressure led schools and teachers to focus resources on students near the cutoff, rather than those far below (below basic) or far above (advanced) the cutoff (Neal & Schanzenbach, in press). These shares vary across states for two reasons: (1) differences in achievement levels of students and (2) differences in the stringency of state tests and standards. The interaction of share near cutoff with an indicator for years after 2003 is added in the regression. In column 4 of Table 1, the variable is included as an exogenous control. In column 5, we account for the endogeneity of 2003 test performance by instrumenting for the share near proficient with an external measure of the stringency of the state test (NCES, 2007). Unfortunately, the measure is not available for all states, and the instrumental variables estimates are only for a subsample of states. In both columns, the impact of NCLB pressure is positive (and significant in column 4), but the estimated FLE and year interactions are largely unchanged.

We conduct additional robustness checks for the main finding by considering other macroeconomic and policy changes during the sample period that may have differentially affected FLE and FLI children. Each change is added in our basic model separately. First, we include interactions between FLE and a wider set of controls for the state economy: unemployment rates for women and for single mothers, as well as employment shares in each of 10 industry categories. Second, we add a control for Medicaid coverage rates of children 10 and under interacted with FLE.[19] Third, we consider the roles of state support for childcare and early childhood education. A component of the reforms we study was the increased availability of funding for states to subsidize childcare expenses to assist single mothers in entering the paid labor force. To determine if this expansion of publicly funded childcare support was a particular source of test score gains, we control for the state-by-year share of the child population (birth to age 14) served by the Child Care and Development Fund (CCDF), which was created under PRWORA and replaced the previous welfare-related child care programs provided under the Social Security Act. We include the measure of CCDF use and its interaction with FLE as additional controls in the basic model. Separately, we control for CCDF using a constant-dollar state-by-year measure of expenditures per child population.[20]

Last, we consider the potential impact of exposure to prekindergarten (pre-K) on test scores. Before and during the sample period, several states introduced or

[18] The data are obtained from each state Department of Education Web page. For a few states without 2003 information, we use the share for 2004.

[19] These are obtained from March CPS files, averaged over the relevant time periods.

[20] The state–year counts of children served by CCDF are from the U.S. Department of health and Human Services Web site (http://www.acf.hhs.gov/programs/ccb/data/index.htm) covering 1998 to 2006. Population data are from the U.S. Census. Data for state–year CCDF spending are from U.S. DHHS and the House Ways and Means Committee 2004 Green Book (online at http://www.gpoaccess.gov/wmprints/green/2004.html) and cover 1995 to 2006. We use the value for the year prior to the NAEP assessment.

expanded their early childhood education programs through publicly funded pre-K. Previous work has found a positive impact of universal pre-K on NAEP scores for children residing in rural areas (Fitzpatrick, 2008). We reestimate our main regression models on several subsamples of children who are unlikely to have benefited from pre-K expansions: children in states with small pre-K programs (enrolment under 5 percent of the 4-year-old population in 2002), children in states without statewide pre-K programs in place by 2002, and children in all non-waiver states who do not reside in rural areas.[21] Regression results for all these robustness tests are reported in Table 5 of the online Appendix.[22] The main results are unchanged in each of these regressions.

To provide a direct link between the treatment effects estimated above and actual welfare policy changes, we calculate cross-state correlations between the size of the treatment effects in 2005, the year in which gains appear to level off, and the intensity of the welfare reforms, measured by the 1996 caseload ratio and the change in caseload ratio between 1996 and 2005. The caseload ratio is defined in each state as the ratio of the number of families receiving cash welfare benefits to the number of children in that state receiving free school lunches. States with higher caseload ratios in 1996 have greater fractions of FLE students likely to be affected by changes in welfare rules, and states with larger caseload declines between 1996 and 2005 are likely to have enforced their welfare policy changes more effectively. We expect the treatment effect to be larger in states where the FLE treatment group experienced more intense treatment from welfare reform. Indeed, for non-waiver states, the correlation between the PRWORA treatment effect and the 1996 caseload ratio is 0.36 (significant at the 5 percent level); the correlation with the *change* in caseload ratio between 1996 and 2005 is -0.50 (significant at the 1 percent level). This reinforces the positive association between welfare reform and relative math performance gains.

Heterogeneous Effects by Race, Sex, and Ability

In this section, we explore the ways in which our estimated impact of welfare reform varies by student characteristics. We first consider demographic variables such as sex and race and next consider the student's position on the test score distribution.

Table 2 reports estimates of the basic model and the fully interactive model with additional interaction terms between race and sex and the treatment indicators. In the first two columns, the time trends are assumed to be common to boys and girls, whites and non-whites. The treatment effects for white females are captured in the FLE*Year terms. These are positive and significant. The positive and significant Male*FLE*Year terms indicate that the gains associated with welfare reform are larger for boys, and the positive and significant NonWhite*FLE*Year terms indicate the gains are also lager for non-whites. In columns 3 and 4 of the table, we relax the assumption of common time trends and include additional interactions between the year indicators and the sex and race indicators, allowing the national trends to differ by race and by sex. This change has little effect on the basic model (column 3 compared to column 1), but reduces the estimated racial difference in treatment effects in the fully interactive model and renders it statistically insignificant (column 4).

The finding of larger treatment effects for non-whites, although not always significant, is consistent with the higher welfare use rates among non-whites than

[21] Data are from the National Institute of Early Education Research (NIEER) Yearbook 2007 (www.nieer.org).
[22] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

whites, even conditional on free lunch eligibility.[23] The finding of larger treatment effects for boys is consistent with the experimental literature and may result from girls spending more time on housework when their mothers participate in the labor market. It suggests that the direct channel whereby welfare reform reduces the incentives for girls to become single mothers is not of primary importance for 4th-grade test scores.

Another dimension over which the impact of welfare reform may vary is student ability. In Table 3, we report estimates of the treatment effects of welfare reform on low-income students at various points along the test score distribution using the Athey and Imbens (2006) “changes-in-changes” (CC) nonlinear difference-in-differences estimator.[24] The CC model relaxes the linearity of the standard DD model and matches quantiles in the FLE and FLI groups for computing counterfactuals based on test scores in the base year. In our context, the CC model allows the test score production function to vary over time, and the distribution of innate ability to vary by income group, but assumes that the distribution of ability is unchanged within each group over time and that the production function (mapping ability to test scores) does not vary with income.

Each column of the table reports estimated effects at a different test score percentile: 10th, 25th, median, 75th, and 90th. Estimates are adjusted for the basic set of controls in column 2 of Table 1. The sample excludes students in early waiver states. Each row presents a separate set of CC estimates for the FLE treatment effect between 1996 and a later year: 2000, 2003, 2005, or 2007. The standard errors in Table 3 are computed using bootstrapping with 1,000 replications. The estimates should be interpreted in the same manner as the main results in column 2 of Table 1. They are qualitatively similar to those in Table 1, as welfare reform is associated with substantial test score gains across the test score distribution. The CC estimates of test score gains level off between 2005 and 2007 across the entire test score distribution. In all years, the CC approach finds greater gains for lower ability students.

Value-Added Estimation

In this section, we address concerns that the treatment effects estimated in the previous sections may in part be attributable to the 1993 and earlier expansions of the earned income tax credit (EITC) that also altered work incentives for low-income families. The EITC expansions have been associated with increased labor supply in low-income families (Hotz & Scholz, 2003) and with improvements in children’s academic achievement (Dahl & Lochner, 2005). These effects are of the same direction as those of the welfare reform; not controlling for EITC may bias the estimated effect of welfare reform upward. The timing of EITC expansions implies that they may affect school performance in our sample through effects on family inputs during early childhood.

Our strategy to address this concern is to use a value-added framework, in which we estimate the test score *growth* between 4th and 8th grades for FLE and FLI groups in different cohorts. We exploit the fact that the 1987 birth year cohort was tested in 1996 as 4th graders and in 2000 as 8th graders; the 1991 cohort was tested in 2000 as 4th graders and could potentially be tested in 2004 as 8th graders;[25] and the 1994 cohort was tested in 2003 as 4th graders and in 2007 as 8th graders. Hence,

[23] For example, 1996 SIPP data show that the rate of welfare use among FLE families was 18.4 percent for whites, 41.7 percent for African Americans, and 28.7 percent for Hispanics. In 2007, welfare use was lower for all groups, but still substantially higher for non-whites: 5.3 percent for whites, 15.5 percent for African Americans, and 8.85 percent for Hispanics.

[24] Estimation was implemented in Matlab using a modified version of the Athey and Imbens code, available for download from Professor Athey’s Web page at http://kuznets.harvard.edu/~athey/.

[25] 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.

we can estimate the performance of 8th graders in 2000, 2004, and 2007 while controlling for the performance of their synthetic cohort 4 years earlier in 4th grade. This removes the effects of confounding factors that are experienced in early childhood or the first four years of schooling, and thus only inputs between 4th and 8th grades affect the 8th grade test score. The model is as follows:

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

where $t = 2000$ for cohorts born in 1987, 2004 for those born in 1991, and 2007 for those born in 1994. Between the 1996 test in 4th grade and the 2000 test in 8th grade, students born in 1987 were exposed to EITC for 4 full years, but to TANF for only 2 to 3 years. The later birth year cohorts were exposed to both EITC and TANF for the full 4 years between 4th and 8th grades. Thus, in our value-added approach we hold EITC exposure constant and assess the impact of 4 full years of exposure to welfare reform on the 1991 and 1994 cohorts relative to the shorter exposure of the 1987 cohort.

We estimate a linear model similar to Equation (2):

$$\begin{aligned} P_{ist}^{g=8} &= P_{is,t-4}^{g=4} \cdot \alpha + \beta_s + \beta_{FLE} \cdot FLE + \beta_t + \tau_t + \beta_{FLE,t} \cdot FLE + \tau_t \\ &\quad + \beta_X \cdot X_{ist} + \nu_{ist} - \nu_{is,t-4} \end{aligned} \tag{3}$$

In Equation (3), the relevant family inputs for the 1987, 1991, and 1994 cohorts are averages between 1996 and 1999, between 2000 and 2003, and between 2003 and 2006, respectively. Estimation results are shown in Table 4. The basic specification is in column 1 and the fully interactive is in column 2. The coefficients on 4th-grade test scores are about 0.5, significantly lower than unity. There are no significant differences in test score growth between 4th and 8th grades for FLI students across the cohorts. However, FLE students experience more growth in later cohorts. In the full model, the 1991 birth cohort of FLE students has 1.8 points greater test score growth, relative to FLI students, than the 1987 birth cohort; the 1994 cohort has 2.5 points greater relative test score growth than the 1987 cohort. These estimates are all significantly different from zero at the 5 percent level, but they are not significantly different from each other. The value-added estimates do not vary significantly by sex, but they do vary by race. The FLE–cohort interactions are nearly twice as large for non-whites as for the overall sample (column 3).

Given the timing of these effects, they are not easily attributable to the EITC expansions, and likely reflect the impact of welfare reform. However, we cannot conclusively rule out EITC as a contributing factor to the observed test score gains, as we implicitly assume in Equation (1') that the impact of 4-year exposure to EITC expansion on relative test score gains is constant between 1996 and 1999, between 2000 and 2003, and between 2003 and 2006. Instead, it may be more appropriate to attribute the test score gains to the broader package of welfare reforms that includes both EITC expansions and changes to AFDC/TANF.

Preexisting Trends in the Achievement Gap

This section addresses the concern that the observed narrowing of the achievement gap by income during our sample period was not caused by welfare reform, but was instead the result of a preexisting trend. Because information regarding student free lunch eligibility was not collected in the NAEP prior to 1996, we approach the preexisting trend indirectly, using student demographics and school level income from the Restricted Use NAEP records for the 1990 and 1992 assessments. Since the 1990 state assessment was only conducted for 8th graders, we use the national sample for that year, which contains only 8,790 observations.

First, we compare the test score trends of students in different racial and ethnic groups. Subsidized lunch eligibility varies greatly by student race and ethnicity: Only 23 percent of white students are eligible for subsidized lunches, as compared to 72 percent of African American students and 69 percent of Hispanic students. However, racial groups provide a much noisier proxy for welfare exposure than FLE. As we report above, in 1996, 25.4 percent of FLE children received welfare payments but virtually no FLI children received welfare payments. Welfare receipt by race is less clear-cut: While greater shares of African American and Hispanic children received welfare payments in 1996 (19.6 and 13.5 percent, respectively), a sizeable percent of white children (3.5) also received welfare payments.[26]

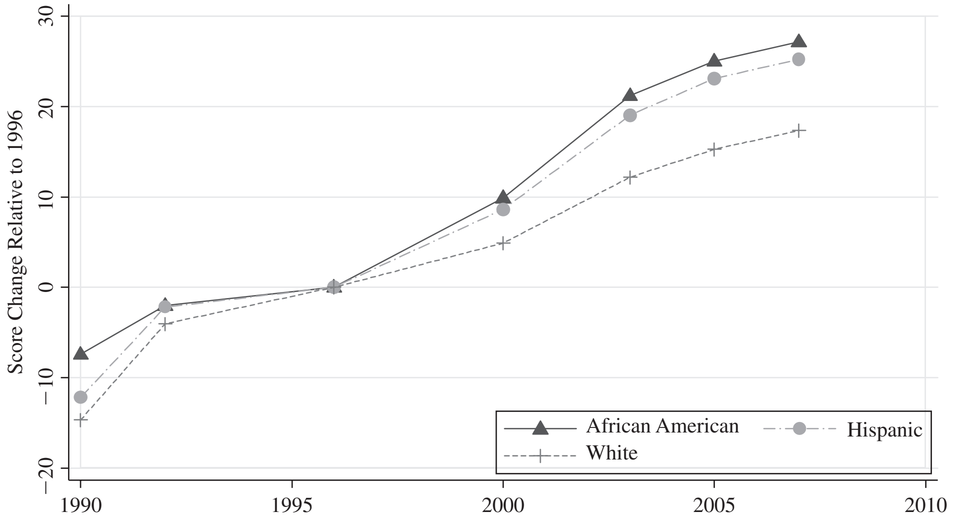
Figure 2A plots the changes in average test scores relative to the 1996 baseline for white, African American, and Hispanic students in non-waiver states. In the years preceding welfare reform (1990 to 1996), test scores of white students improved relative to those of Hispanic and African American students. However, after 1996, African American and Hispanic students experienced relative test score gains. This pattern confirms the post-reform gains, but does not support a preexisting narrowing of the test score gap by income. In a related DD regression with year indicators, race indicators, and their interactions as the sole covariates, we test the significance of these relative trends.[27] Between 1990 and 1996, African Americans experienced significantly less test score growth than whites. For Hispanics between 1990 and 1996, and for both African Americans and Hispanics between 1992 and 1996, test score growth was statistically equivalent to that of whites. Each of the DD estimates after 1996 is positive and significant, indicating relative test score improvements for African American and Hispanic students following national welfare reform. As above, the gains are significant by 2000, prior to the passage of NCLB. Further support for the timing is suggested from the 11 early waiver states. Results from a similar regression on the early waiver states show no relative gains for African American or Hispanic students between 1990 and 1992 but significant relative gains after 1992 (by 1996 for African Americans and by 2000 for Hispanics).

In our second test for preexisting trends, we use school-level income information. Although individual student income is not available before 1996, the NAEP does collect school level income information on subsidized lunch participation (rather than eligibility) in two earlier waves (1990 and 1992). We compile a measure of school poverty rates for each NAEP student, based on the share of students in their school who participated in the subsidized lunch program (SLP), a subset of SLE. We group schools according to their SLP shares in three broad categories: very low share SLP with under 10 percent of students SLP, medium share between 10 and 50 percent, and high share over 50 percent. One caveat regarding these data is their small sample sizes for 1990 and 1992. As mentioned above, the 1990 assessment has fewer than 9,000 observations. In addition, for 1992, the school income question was only asked in the school supplemental survey, leading to only 4,031 observations.

Figure 2B plots test score deviations from 1996 values for non-waiver states by school income category. The figure shows a widening of the test score gap between higher- and lower-income schools prior to 1996 and a narrowing between 1996 and 2007. This suggests a break from trend around the time of welfare reform. In related regression analysis, we test the significance of these relative trends. From 1990 to

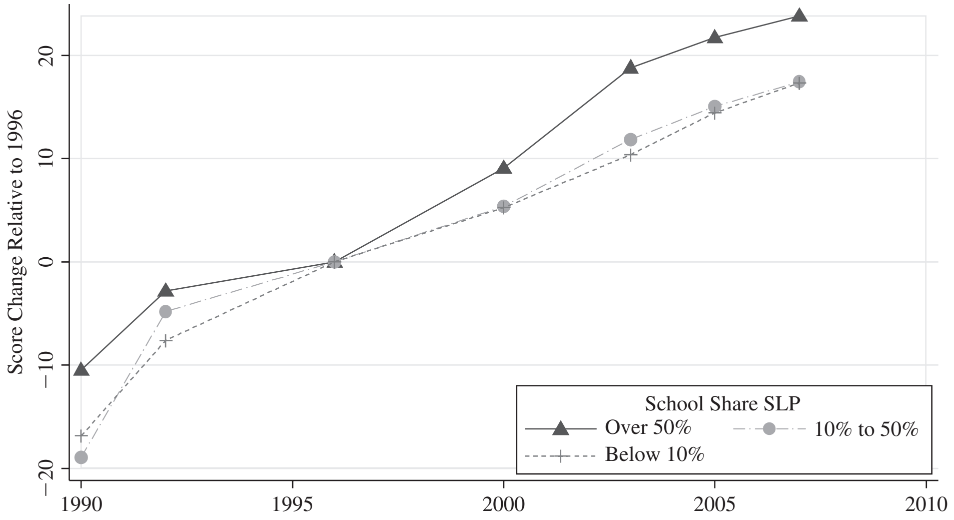
[26] There is also more heterogeneity in income within racial groups than within the FLE group. For example, the average 1996 incomes of African American and Hispanic households with children ($2,461 and $2,500, respectively) are much higher than the average income of FLE group ($667) and have much larger within-group variation. Thus, the DD estimate of the treatment effect of welfare reform using African American and Hispanic as the treatment group and white as the control group will be more attenuated (authors' calculation from SIPP Wave 1 of 1996 Panel).

[27] Standard errors are clustered at the state level.



Source: NAEP Restricted Use Mathematics Assessments: 1990, 1992,2000, 2003, 2005, and 2007. Figure plots test score deviations from the 1996 value within each demographic category. Sample consists of non-waiver states.

Figure 2A. 4th-Grade Math Score Deviations from 1996 by Race and Ethnicity.



Source: NAEP Restricted Use Mathematics Assessments: 1990, 1992, 2000, 2003, 2005, and 2007. Figure plots test score deviations from the 1996 value within each school share SLP (subsidized lunch participation) category. Sample consists of non-waiver states.

Figure 2B. 4th-Grade Math Score Deviations from 1996 by School Share SLP.

1996 the DD estimates for the widening of the test score gap between schools with high and low shares SLP is statistically insignificant. The relative narrowing following welfare reform is significant at the 5 percent level from 2000 through 2007. Turning to the early wavier states, we estimate DD treatment effects relative to the base year of 1992. For low-income students in these states, there is a relative widening of the test score gap between 1990 and 1992 and a relative narrowing that starts in 1992. Hence, for waiver and non-waiver states, the time pattern of the test score gap is consistent with a break from trend around the time of welfare reform.

The analysis in this section detects no preexisting trends in the test score gaps by race or by school income in the years leading up to welfare reform. This provides some support for the validity of the main results by linking the timing of the relative gains to welfare reform. The time patterns of the relative gains by sex and by school income category are also consistent with the evidence on educational attainment: Miller and Zhang (2008a) find relative gains in educational attainment for low-income children in the years following state welfare reforms, but no relative gains in the preceding years.

EFFECTS OF STATE WELFARE REFORMS

We exploit variation in test scores before and after *national* welfare reform to establish the main results. This section tests the robustness of the main results to the inclusion of all states and the use of an identification strategy that exploits both cross-state and over-time variation in welfare policies. In addition to estimating the average impact of any welfare reform, we attempt to disentangle the effects of various welfare policy changes.

We estimate the differential relationship between test scores and time exposed to any statewide reform for lower- and higher-income students as modeled in Equation (4):

$$\begin{aligned} P_{ist} = \beta_s + \beta_{FLE} \cdot FLE_i + \beta_t \cdot \tau_t + \beta_{FLEMWR} \cdot FLE_i \cdot MWR_{st} \\ + \beta_{MWR} \cdot MWR_{st} + \beta_X \cdot X_{ist} + \nu_{ist}, \end{aligned} \tag{4}$$

where our variable of interest is the interaction term between MWR_{st} (months of exposure to welfare reform) and free lunch eligibility. The treatment variable captures the differential trend in test scores between FLE and FLI students as an increasing function of their duration of exposure to welfare reform. The variable MWR_{st} is equal to the number of months between the state welfare waiver or TANF implementation, up to a maximum of 108 months (representing the age of a typical 4th grader). As discussed above, the impact of welfare reform may increase with months of exposure because of a direct duration effect or an indirect effect related to age at first exposure. Although the impact of welfare reform may not increase linearly, our limited sample precludes a more complete investigation of the underlying nonlinearities.

Table 5 reports the main estimates from Equation (4). In columns 1 and 2, the FLE*MWR interaction is estimated at 0.076 (significant at 1 percent) in both the basic model and the fully interactive model. The point estimate of 0.076 per month implies total gains of 10 test points between 1996 and 2007, slightly larger than the 9 test points gain in the comparable DD model for non-waiver states

[28] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

Table 5. Welfare reform timing and 4th-grade math scores.

Model	Basic 1	Full 2	Basic 3	Full 4
FLE*Months exposure to reform	0.076**	0.076**		
	(0.010)	(0.010)		
Months exposure to reform	−0.013			
	(0.022)			
FLE*State highest income tax rate			−0.189	−0.141
			(0.139)	(0.155)
State highest income tax rate			−0.290	
			(0.290)	
Observations	703,530	703,530	703,530	703,530

Notes: All regressions use Restricted Use Data for the full sample of states. Other notes are as in Table 1.

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

(columns 2 and 3 of Table 1). Full results for control variables are in Table 4 of the online Appendix.[28] The MWR interaction for the slightly higher-income RLE group is also positive, but substantially smaller and not significantly different from zero (not reported). This falsification check confirms that we fail to detect an impact of welfare reform for RLE students.

The last two columns of Table 5 report a second falsification check based on estimating the impact of policy variation that occurred during our sample period, but that is not expected to lead to test score gains for low-income students. We selected changes in the top marginal tax rate for state personal income taxes.[29] During the sample period, 27 states changed their top marginal rates (16 decreases, 8 increases, and 3 combinations). We estimate the effect of current tax rates on child test scores. When the top tax rates are increased, the highest-income families in the state pay a greater share of their income in state taxes. The potential direct impact of the rate increases will apply to only a small subset of students in the sample, and the indirect effects will be negligible in our regression model that controls for educational spending. Indeed, the estimated treatment effects in the basic model (column 3) and the fully interactive model (column 4) are negative and statistically insignificant. This lack of an effect provides some reassurance that our empirical approach is able to reject the impact of policies with no relation to relative test scores.

We next investigate the effects of three specific welfare policies: time limits, sanctions, and school requirements for dependent children.[30] The effect of each specific policy is estimated using Equation (4), but with MWR_{st} replaced by the months of exposure to the specific policy element. The estimation results from the fully interactive models are reported in Table 6. For each additional month of exposure to a state time-limit policy, sanction policy, or school requirement policy, FLE students experienced relative test score gains of 0.046, 0.08, and 0.041 points. When all three policies are included along with the general state waiver policy, which can be interpreted as the collection of all welfare policies, only the school requirement policy is significant at the 10 percent level. The lack of precision in the estimated effects of particular

[29] We thank an anonymous referee for this suggestion.
[30] In separate analysis, we explore the effects of other specific welfare policies—diversion payments, job search requirements at application, health screening, and regular check-up requirements for dependent children, but none of these has any effect on test score gains, individually or jointly with other policies. We also explore the impact of stricter time limit and sanction policies, because they may provide different incentives for current and potential welfare recipients. The estimation results, not reported, find no evidence that increasing the severity of welfare policies generates additional test score gains.

Table 6. Elements of welfare policy reform and 4th-grade math scores.

Model	Full 1	Full 2	Full 3	Full 4	Full 5
FLE*Months exposure to reform	0.076**				0.038
	(0.010)				(0.034)
FLE*Months any time limits		0.046**			0.013
		(0.010)			(0.010)
FLE*Months any sanctions			0.08**		0.012
			(0.012)		(0.038)
FLE*Months school req.				0.041**	0.02+
				(0.008)	(0.010)
Observations	703,530	703,530	703,530	703,530	703,530
P-value on test of joint significance of FLE*reform variables:					0.000

Notes: All regressions use Restricted Use Data on all states. Regressions are estimated using the Full Model, corresponding to column 3 in Table 1. Months of exposure to reform counts the number of months of the student's life from the earliest statewide waiver or TANF implementation to the NAEP test. Any time limit includes lifetime or spell-specific time limits. Any sanctions include full, partial and gradual sanctions. School requirements link welfare receipt to child schooling. See text for details. Other notes as in Table 1.

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

policies may result either from the collinearity of the policy measures or from measurement error. Nevertheless, the welfare policies jointly matter: An *F*-test of the joint significance of waivers and specific policies rejects zero at the 1 percent level.

CONCLUSION

This paper presents the first analysis of the impact of U.S. federal and statewide welfare reforms in the 1990s on the academic achievement of children in low-income families. We estimate the net effect of welfare reform in a reduced form, difference-in-differences framework, using nationally representative mathematics test score data. Children in higher-income families are used as a control group, and changes in their test scores provide a counterfactual for what would have happened to low-income students absent welfare reform. We find no evidence that welfare reform harmed the academic performance of low-income students. On the contrary, welfare reform is associated with a substantial narrowing of the mathematics test score gap between high-income and low-income students in 4th and 8th grades. These benefits are present for low-income students irrespective of race, gender, and innate ability. The finding of relative test score gains for low-income students is robust to controlling for a wide range of observable and unobservable factors, including state macroeconomic changes, and variation in educational inputs and policies.

The relative test score gains for low-income children is statistically and economically significant starting in 2000, the first test year following national reform and only 4 years after the passage of PRWORA. These short-term educational gains are consistent with the results from earlier studies of welfare reform experiments in both sign and magnitude, and support the external validity of the experiments to a nationwide sample experiencing the actual statewide and national reforms.

This paper extends the time horizon of analysis to more than a decade following welfare reform, presenting new evidence of long-term gains. The test score gains increase in size in the years following welfare reform, leveling off in 2005 for 4th graders. Although the long-term gains after 2000 are crucial for policy evaluation, the later estimates are more susceptible to outside confounding factors, including economic and educational policy changes. The role of time-varying omitted factors, as indicated by test score changes for students in the FLI control group, is small and insignificant between 1996 and 2000, but growing in importance from 2003 to 2007. As a result, the longer-term gains may be less reliably estimated than the short-term gains. Nevertheless, our estimates of a continued narrowing of the test score gap between 2000 and 2005 provide evidence for persistent benefits from welfare reform over a 10-year horizon.

The reduced-form approach in this paper, however, is unable to pinpoint a single underlying mechanism responsible for welfare reform's beneficial effects. Welfare reform led to multiple behavioral changes in low-income parents, including increased labor force participation and reduced welfare receipt. A thorough exploration of the channels for the test score gains is a promising avenue for future research.

AMALIA R. MILLER is assistant professor of Economics, University of Virginia.

LEI ZHANG is assistant professor of Economics, Clemson University.

ACKNOWLEDGMENTS

We are grateful for comments from four anonymous referees, Doug Bernheim, William Dougan, Gabrielle Fack, Eric Hanushek, Caroline Hoxby, John Pepper, Macke Raymond, Sarah Turner, and Ebonya Washington; seminar participants at Clemson, Maryland, San Francisco State, Stanford, UC–Merced, Virginia, and the BLS; and conference participants at the AEFA, Econometric Society, NBER Education Program, NBER TAPES, SEA, and SOLE. Liliana Danila, Felicia Emery, and Maria Fitzpatrick provided excellent research assistance. Financial support was provided by the University of Virginia Bankard Fund for Political Economy and Clemson University College of Business and Behavioral Science.

REFERENCES

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

Baum, C. L. (2003). Does early maternal employment harm child development? An analysis of the potential benefits of leave taking. Journal of Labor Economics, 21, 409–448.

Blank, R. M. (2002). Evaluating welfare reform in the United States. Journal of Economic Literature, 40, 1105–1166.

Blau, D. M. (1999). The effect of income on child development. Review of Economics and Statistics, 81, 261–76.

Blau, F. D., & Grossberg, A. J. (1992). Maternal labor supply and children's cognitive development. Review of Economics and Statistics, 74, 474–81.

Bleakley, H. (2007). Malaria eradication in the Americas: A retrospective analysis of childhood exposure. Mimeo.

Crouse, G. (1999). State implementation of major changes to welfare policies, 1992–1998. Available at http://aspe.hhs.gov/hsp/Waiver-Policies99/policy_CEA.htm.

Dahl, G., & Lochner, L. (2005). The impact of family income on child achievement. Working Paper 11279, National Bureau of Economic Research.

Duncan, G. J., & Chase-Lansdale, P. L. (2001). Welfare reform and children's well-being. In R. M. Blank & R. Haskins (Eds.), The new world of welfare (pp. 391–412). Washington, DC: Brookings Institution Press.

Fitzpatrick, M. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. B.E. Journal of Economic Analysis & Policy (Advances), 8(1), Article 46.

Fletcher, S. H., & Raymond, M. E. (2002). The future of California's academic performance index. CREDO, Hoover Institution, Stanford University. Mimeo.

Gennetian, L., Duncan, G., Knox, V., Vargas, W., Clark-Kauffman, E., & London, A. (2004). How welfare policies can affect adolescents: A synthesis of evidence from experimental studies. Journal of Research on Adolescence, 14, 399–423.

Goertz, M. E., & Duffy, M. C. (2001). Assessment and accountability systems in the 50 states: 1999–2000. Consortium for Policy Research in Education, Graduate School of Education, University of Pennsylvania. RR-046.

Gregg, P., Washbrook, E., Propper, C., & Burgess, S. (2005), The effects of mother's return to work decision on child development in the UK. Economic Journal, 115, F48–F80.

Grogger, J., & Karoly, L. A. (2005) Welfare reform: Effect of a decade of change. Cambridge, MA: Harvard University Press.

Hanushek, E. A. (2002). Publicly provided education. In A. J. Auerbach & M. Feldstein (Eds.), Handbook of public economics (pp. 2045–2141). Amsterdam: North-Holland.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? Journal of Policy Analysis and Management, 24, 297–327.

Hanushek, E. A., & Zhang, L. (2006). Quality-consistent estimates of international returns to skill. Working Paper 12664, National Bureau of Economic Research.

Haskins, R. (2001). Effects of welfare reform on family income and poverty. In R. M. Blank and R. Haskins (Eds.), The new world of welfare (pp. 103–136). Washington, DC: Brookings Institution Press.

Hotz, V. J., & Scholz, J. K. (2003). The earned income tax credit. In R. A. Moffitt (Eds.) Means-tested transfer programs in the United States (pp. 141–197). Chicago: University of Chicago Press.

Krueger, A. B. (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, 114, 497–532.

Miller, A. R., & Zhang, L. (2008a). Intergenerational effects of welfare reform. Mimeo.

Miller, A. R., & Zhang, L. (2008b). The effects of welfare reform on the academic performance of children in low-income households. Mimeo.

Moffitt, R. A. (2003). The temporary assistance for needy families program. In R. A. Moffit (Eds.), Means-tested transfer programs in the United States (pp. 291–363). Chicago: University of Chicago Press.

Morris, P., Duncan, G., & Clark-Kauffman, E. (2005). Child well-being in an era of welfare reform: The sensitivity of transition in development to policy change. Developmental Psychology, 41, 919–932.

Murnane, R. J., Willet, J. B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. Review of Economics and Statistics, 77, 251–266.

National Center for Educational Statistics (NCES). (2007). Mapping 2005 state proficiency standards onto the NAEP scales. NCES 2007–482.

Neal, D., & Schanzenbach, D. W. (in press). Left behind by design: proficiency counts and test-based accountability. Review of Economics and Statistics.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73, 417–458.

Ruhm, C. J. (2004). Parental employment and child cognitive development. Journal of Human Resources, 39, 155–192.

U.S. Department of Health and Human Services (USDHHS). (1997). Setting the baseline: A report on state welfare waivers. Available at http://aspe.hhs.gov/hsp/isp/waiver2/title.htm.

Waldfogel, J. (2007). Welfare reforms and child well-being in the US and UK. Working paper, Centre for Analysis of Social Exclusion.