Mansfield, Jonathan; Slichter, David

**Working Paper**

# The Long-Run Effects of Consequential School Accountability

IZA Discussion Papers, No. 14503

**Provided in Cooperation with:**
IZA – Institute of Labor Economics

*Suggested Citation:* Mansfield, Jonathan; Slichter, David (2021) : The Long-Run Effects of Consequential School Accountability, IZA Discussion Papers, No. 14503, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/245554

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14503

# The Long-Run Effects of Consequential School Accountability

Jonathan Mansfield
David Slichter

JUNE 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14503

# The Long-Run Effects of Consequential School Accountability

**Jonathan Mansfield**
*Binghamton University*

**David Slichter**
*Binghamton University and IZA*

JUNE 2021



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## The Long-Run Effects of Consequential School Accountability*

The rise of high-stakes accountability programs was one of the most noticeable changes in the U.S. education system during the 1990s and early 2000s. We measure the impact of these programs on students' long-run outcomes. We find that exposure to accountability modestly but detectably increased educational attainment – roughly .02 years per year of exposure. Effects on income were positive, but again modest and insignificant in most specifications. Lastly, if accountability had substantial effects on human capital, treated individuals would be expected to sort into occupations requiring greater use of tested (math and literacy) skills, potentially at the expense of non-tested skills. Instead, we find that accountability had virtually no effect on occupational requirements. Our results suggest that accountability was likely net beneficial for students' long-run outcomes, but not transformative.



**JEL Classification:** I28, J24, H0

**Keywords:** accountability, long-run effects, teacher incentives, teaching to the test



**Corresponding author:**
David Slichter
Binghamton University
4400 Vestal Pkwy E
Binghamton
NY 13902
USA

E-mail: slichter@binghamton.edu



* We are grateful for comments from Greg Caetano, Sulagna Mookerjee, and Sol Polachek. All mistakes are ours.

# 1 Introduction

During the 1980s, many policymakers in the United States grew concerned about the quality of public schools, particularly following the publication of *A Nation at Risk* (National Commission on Excellence in Education 1983). Over the following years, and especially during the 1990s, many states responded by implementing laws aimed at making the public school system accountable to the public. The period of rapid growth in accountability programs culminated with the No Child Left Behind (NCLB) Act of 2001, a law which required any state receiving federal education funding to implement an accountability system – a sufficiently persuasive incentive that all remaining states adopted accountability.

Accountability programs have taken several forms. In some cases, accountability programs simply involved releasing information on standardized test score performance to the public, often in the form of a school report card. In other cases, accountability was *consequential*: Schools and/or teachers were given explicit incentives to meet some standard of test score achievement for their students, ranging from financial incentives (e.g., receiving a payment for high test score performance) to the threat that low-performing schools would be closed.

Accountability programs have been controversial since their inception.

Supporters argue that accountability programs promote transparency, give beneficial incentives to educators, and might help weed out dysfunctional teachers and schools (Hanushek and Raymond 2001, McKenzie and Kress 2015). There is indeed strong evidence that accountability programs increase test score performance on the test used to evaluate students (e.g., Richards and Sheu 1992, Ladd 1999, Greene 2001) – though the evidence for effects on low-stakes exam performance is more mixed, with some evidence pointing to increases and other evidence pointing to no effects; see Figlio and Loeb (2011) or the National Research Council (2011) for detailed reviews. Positive effect estimates often indicate that accountability increases low-stakes test performance by on the order of .1 standard deviations (e.g., Lee 2008, Dee and Jacob 2011, NRC 2011, Wong et al. 2015). There is also some evidence that accountability programs have effects on personnel quality (e.g., Clotfelter et al. 2004, Loeb and Cunha 2007, Feng et al. 2018). With respect to transparency, while school report cards may affect home prices (Figlio and Lucas 2004), effects on test scores appear to be concentrated only among accountability programs which are consequential (Hanushek and Raymond 2005).

Opponents of accountability programs have several concerns. One is the possibility of "teaching to the test": Educators may alter instruction to improve test scores at the expense of aspects of education quality which are not measured by standardized tests. Educators report decreased instructional time spent on non-tested subjects (Shepard and Dougherty 1991, Stecher et al. 1998, Dee et al. 2013), and the ambiguous evidence about performance on low-stakes tests plus evidence that accountability pressure increases cheating and gaming (Jacob and Levitt 2003, Figlio and Winicki 2005, Jacob 2005) point to the possibility that test score gains from accountability programs may not reflect deep understanding. There are also concerns that accountability and

the accompanying pervasive standardized testing might increase anxiety for students and teachers (Barksdale-Ladd and Thomas 2000, Hoffman et al. 2001), potentially driving away talented teachers or reducing students' intrinsic interest in learning.

In this paper, we estimate the effects of early consequential accountability programs on students' long-run outcomes. We have two research questions. First, what was the overall effect of accountability on students' educational attainment and earnings in adulthood? Second, is there evidence that accountability programs led students to possess skills in adulthood which are narrowly focused in the tested subjects, at the expense of other skills, as would be predicted if teaching to the test were important?

We answer our research questions using panel methods, exploiting the staggered implementation of accountability policies. We measure skills by combining information about workers' occupations with measures of occupational skill requirements from O*NET, on the theory that e.g. engineers are likely to have stronger math skills while journalists are likely to have stronger verbal skills (Roy 1951).

An important challenge for the application of panel methods is that estimates can be biased in contexts with staggered implementation when treatment has delayed effects (see e.g. de Chaisemartin and D'Haultfœuille 2020, Goodman-Bacon 2021). In our context, implementation was staggered, and there are many reasons why accountability programs might have delayed effects, including that (i) sanctions are typically triggered only after repeated performance failures, (ii) educators may learn over time how to respond to accountability pressure, and (iii) accountability affects the quality of education by changing the quality of personnel, which is inevitably a gradual process. We therefore use empirical methods which we argue can avoid this concern. However, addressing this concern requires us to use sample years where there are a substantial number of observations without exposure to accountability. For this reason, we focus on estimating the impacts of early (in general, pre-NCLB) consequential accountability.

We find evidence of effects of accountability on educational attainment: Each year of exposure to accountability increased average attainment by roughly .02 years.

Estimated effects on income are also positive but less precise. We find that each year of exposure increased income by between .2 and .5%, depending on specification. However, these estimates are mostly not statistically significant.

Lastly, we find zero or near-zero impacts on occupational use of skills expected to be produced by – or crowded out by – instruction aimed at standardized tests emphasizing literacy and numeracy. Occupational use of literacy skills increased by .001-.005 standard deviations per year of exposure, which is significant in some specifications. Use of math skills increased by 0-.001 standard deviations, which is not significant. Effects on the use of creativity, critical thinking, science, and other non-tested subjects are also at most a few thousandths of a standard deviation, have mixed sign, and are almost never statistically significant.

Theoretically, accountability could have had modest overall effects on income and education due to strong but countervailing effects of increased teacher effort and teaching to the test. But

our skill usage estimates suggest that the modest overall effect is because these two effects were each modest, not because they were canceling.

Collectively, our results suggest that accountability was likely net beneficial to students' long-run outcomes. However, despite its reputation as one of the most high-profile (and controversial) education reforms on the last few decades, accountability appears not to have had a transformative effect on any aspect of human capital in adulthood.

Three other papers study the effects of accountability pressures on labor market outcomes. Using data from Texas, Deming et al. (2016) study the effect of experiencing greater vs. weaker accountability pressure among schools within the same accountability regime. They find mixed results, with accountability pressure increasing incomes at schools with low-performing students while decreasing incomes at schools with high-performing students. At very low-performing schools in Israel, Lavy (2020) finds large long-run benefits from a teacher performance incentive, with earnings at ages 28-30 increasing by 8 to 9 percent, though zero effects on income cannot be rejected at the 5% significance level. Relative to these papers, we make two contributions. First, we study the effect of the presence of consequential accountability programs as opposed to their absence, while Deming et al. study variation within accountability programs and Lavy studies a particular performance incentive which is stronger than would be experienced within a typical consequential accountability program. We view these approaches as complementary; focusing on subsamples with unusually strong pressure to increase test scores can increase statistical power to detect the sign of effects, while our approach has the benefit of more realism in understanding whether accountability had appreciable impacts on society as a whole. Second, the set of programs we study is different.[1] Lastly, our use of occupational skill measures allows us to assess whether there are notable effects on skills in a pattern which would be expected to correspond to the effects of accountability pressure.

Finally, Wong (2008) studies the effects of accountability on income, employment, and education using panel methods, finding mixed results. Among various key differences, due to greater data availability, our estimates have an order of magnitude greater precision.[2]

The rest of the paper proceeds as follows. In Section 2, we describe our data sources and institutional background. In Section 3, we describe our econometric methods. Section 4 describes the results. Section 5 concludes.

## 2 Data

Our primary source of data is the American Community Survey (ACS). We use survey years from 2005 to 2017. The key variables from the ACS include labor income, occupation, state of birth,

[1] Deming et al. study an accountability policy in Texas, which is in our sample, but is only one of many accountability programs in our sample.

[2] Other important differences include that our definition of treatment more closely matches actual exposure to accountability, as we discuss in Section 4; various aspects of our specifications; that we use occupational outcome variables; and that we report aggregate effects.

and year and quarter of birth. We use some sample restrictions on birth year and quarter in ways which are motivated and described in Section 3.

**Constructing accountability variables** We assign measures of exposure to accountability to individuals in the ACS on the basis of their state of birth, year of birth, and quarter of birth using the following process.

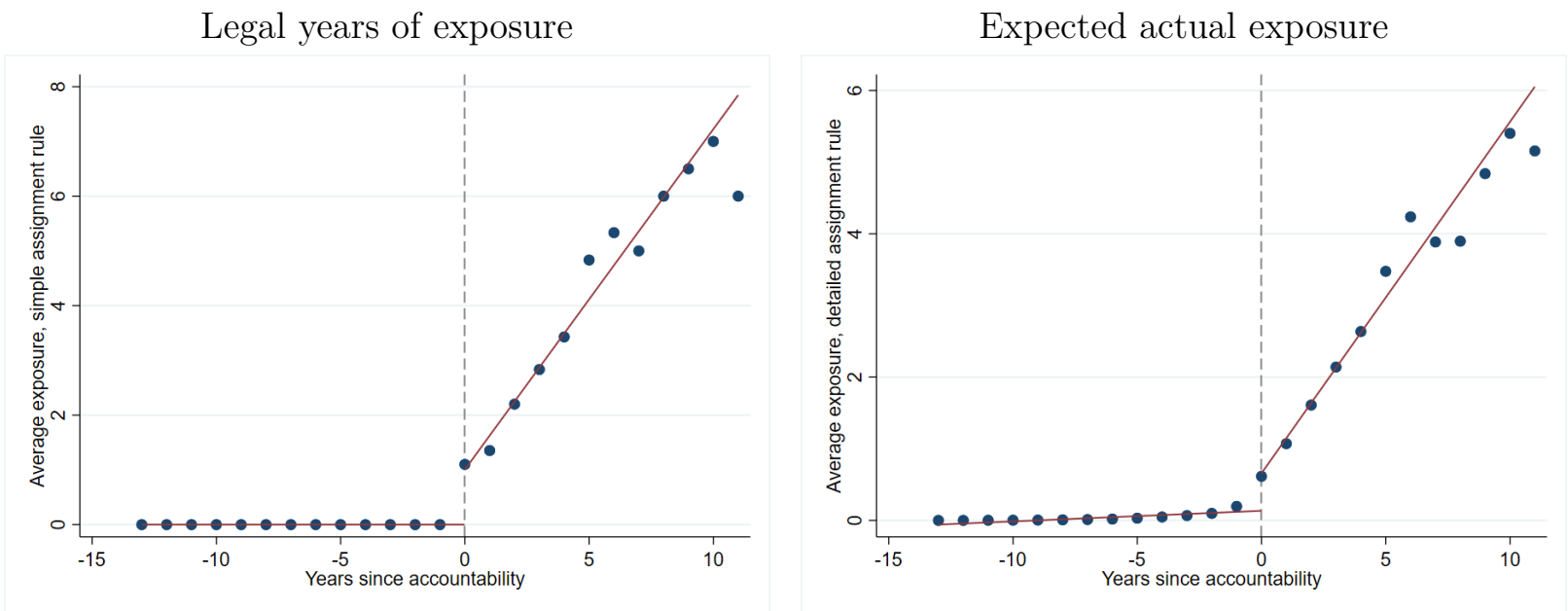
First, we construct the series of pre-NCLB consequential accountability laws. We define an accountability law to be "consequential" if there are either substantial sanctions such as the risk of a school closure or educators being fired, or incentives such as bonus payments. We exclude report card accountability from our measure on the principle that such accountability programs had little effect on contemporaneous outcomes (Carnoy and Loeb 2002, Hanushek and Raymond 2005, Jacob 2005, Dee and Jacob 2011) and therefore presumably generated neither beneficial nor perverse incentives for educators.

Appendix A gives a brief description of the timing of consequential accountability reforms, and grades affected, for each state. For a handful of states, we judged the start of consequential accountability to be ambiguous, due to gradual implementation or the implementation of widespread sub-state accountability. We drop these states from our main analysis, up to an exception described below. In other cases, we judge that, while technically there were incentives attached to test score performance, the accountability regime was so weak (or weakly enforced) that it did not provide a credible incentive for educators. For our main results, these cases are coded as not having consequential accountability, though the results are unaffected by dropping these states from the analysis.

Using the series of laws, we are able to construct the number of years of exposure to accountability which would be experienced by members of a given school cohort who are enrolled in public schools in a particular state. By school cohorts, we mean the collection of students who begin 1st grade in a particular year (e.g., 1991) and progress through school without skipping or repeating grades. By exposure to accountability, we mean that consequences were attached to the test score performance of children in that students' grade. For example, students whose educators had an incentive attached to their standardized test performance in 4th and 5th grades would be considered to have two years of exposure to accountability.

The left panel of Figure 1 shows how the number of years of accountability evolved across school cohorts before and after the start of accountability for our primary sample. The horizontal axis is the number of years before or after the start of accountability in a particular state. So, for example, if the first school cohort which experienced accountability in Vermont were the students who began 1st grade in 1987, then the cohort which began 1st grade in 1986 would be at $-1$ on the horizontal axis, while the cohort which began 1st grade in 1989 would be at 2. The vertical axis shows the average number of years of accountability that the corresponding cohorts would be exposed to. This graph is restricted to the cohorts used in our main analysis; note that this produces an "unbalanced" graph in the sense that not all states have cohorts 10 years after

Figure 1: Accountability by cohort




Years of accountability by cohort. Left: average number of years of exposure to accountability by the number of cohorts since the first cohort within a state to be exposed to accountability. Right: average number of years of expected exposure to accountability (the "detailed" measure) by years before/after the start of accountability in a given state. Averages are taken over state-cohorts.

accountability or 15 years before accountability in our primary sample.

The pattern looks roughly like two lines. By definition, cohorts prior to the start of accountability had no accountability. After the start of accountability, there is a roughly linear increase in the number of years of exposure to accountability. To illustrate why this is the case, consider a state which implements accountability in grades 3-8 but did not previously have accountability. The first affected cohort are 8th graders in the first year of the new policy, and receive one year of exposure to accountability. The second affected cohort gets accountability in both grades 7 and 8, and so on. The relationship is not exactly linear due to gaps in grades affected by accountability (e.g., a state might implement accountability only in grades 4 and 8), changes in the set of affected grades after initial implementation (which in a few cases leads to decreases in accountability by cohort), changes in the set of states represented by each point (because the graph is "unbalanced" in the sense described above), and limits on the number of total grades in which there is accountability.

The ACS data does not contain detailed information on where individuals lived in childhood, what year they began school, or what kind of school they attended. Therefore, we do not know exactly how many years of exposure to accountability each individual in our data had. Instead, we must estimate the number of years of exposure to accountability on the basis of individuals' birth year, quarter, and birth state.

We use two procedures for this assignment.

In the first, which we will call the *simple* assignment, we assign each student to a number of years of accountability that they would receive if they were part of their most likely school cohort

and state of residence. We assume that the student was raised in whichever state they were born in and attended public schools. Additionally, we assume that the student completed first grade in the year equal to the year of their birth plus seven if they were born in January through September, or plus eight if they were born in October through December (e.g., someone born in July of 1982 is assumed to attend first grade in the 1988-89 school year). One complication is that the ACS data report age at the time of the survey and quarter of birth, but do not report the date of the survey, resulting in ambiguity about the respondents' year of birth. Because ACS data are collected year-round, we assume that the year of birth is the survey year minus age for people born in January through June, and equal to the survey year minus (age + 1) for people born in July through December.

In the second, which we will call the *detailed* assignment, we estimate an expected number of years of accountability from birth year, quarter, and birth state. We do this by adjusting for four reasons why the simple assignment rule might be incorrect: (i) students might attend private schools, (ii) students might not attend school in the same state where they were born, (iii) students might not have been in the school cohort that is expected based on their birth year and quarter, and (iv) students might have left school prior to grades in which they would be exposed to accountability.

Appendix B describes our procedures for the construction of the detailed assignment measure. Briefly, it relies on estimation of conditional probabilities of being in private school given grade, which we estimate using 1990 and 2000 Census 5% samples in order to capture probabilities which prevailed at the time our key cohorts were in school; the probability of being in a given grade at a given age given when you were born, which we estimate from the 1980 5% Census sample to overcome limitations in later data; the probability of living in a given state at a given age, conditional on having been born in a particular state, which we estimate using 1990 and 2000 5% Census data; and the probability of still being in school at a given age, which we estimate using the 1990 and 2000 5% samples. As described above, we exclude observations with birth states where the timing of implementing accountability was ambiguous; however, because people from all birth states have a positive probability of moving to these states, we must take a stance on the timing of accountability reforms in these states for the sake of measuring average years of accountability faced by movers. See Appendix B for details.

The right half of Figure 1 illustrates how accountability pressure increases over time under the detailed assignment measure. To construct this graph, we first calculate the average of the detailed measure among all observations with the same birth state and simple-measure cohort. Then, as in the left panel of Figure 1, we assign each such state-cohort to a number of years before or after the start of accountability and take the average among state-cohorts with the same number of years post/prior to the start of accountability.

The graph shows some deviations from the legal rules shown in the left half of Figure 1. In particular, a few students experienced accountability even though the simple measure would not assign them accountability, either because they moved or because they were in a later cohort than

the simple assignment rule would claim. Furthermore, actual exposure to accountability is lower than in the left half of Figure 1 among those assigned to accountability by the simple rule, due to a combination of moving, attending private schools, being in a different cohort than expected based on the simple assignment rule, and leaving school early.

**Measuring skills** We use an individual's occupation in adulthood as a proxy for the skills which they possess. For example, if someone works as a novelist, we might infer that they are good at written communication. The rationales for this approach are that (i) people tend to sort into occupations which match their comparative advantage in skills (Roy 1951), (ii) one of the primary goals of accountability programs is to produce a skilled workforce, meaning that measures of skill which are related to work activities are particularly germane, and (iii) occupation is a convenient measure which is available in large datasets.

In order to translate occupation into a measure of skills, we merge ACS data with data from the Occupational Information Network (O*NET) describing the skill requirements of occupations. O*NET is a resource developed by the United States Department of Labor to help job-seekers find occupations which match their skills and interests. O*NET assigns numerical values to the skill requirements of every US Standard Occupational Classification (SOC) occupation.

We construct skill indices from collections of O*NET's skill measures using the following process. First, we normalize each individual skill measure used in the construction of the index. Next, we take the sum of these normalized measures. Finally, we normalize the sum.

The indices we construct are as follows, using measures listed under the O*NET categories Skills, Knowledge, and Abilities.[3] We construct a "math" index out of the measures Mathematics (Skills), Mathematical Reasoning (Abilities), and Number Facility (Abilities). We construct a "writing" index from Reading Comprehension (Skills), Writing (Skills), Written Comprehension (Abilities), and Written Expression (Abilities). We construct "creativity" from Originality (Abilities) and Fluency of Ideas (Abilities). We construct "critical thinking" from Critical Thinking (Skills), Judgment and Decision Making (Skills), Operations Analysis (Skills), Systems Analysis (Skills), Deductive Reasoning (Abilities), and Inductive Reasoning (Abilities). We construct "science" from Science (Skills), Biology (Knowledge), Chemistry (Knowledge), and Physics (Knowledge). Finally, we construct an index of "non-tested" skills – designed to capture other subjects which are taught in schools but might be crowded out by teaching to the test – from Fine Arts (Knowledge), Foreign Language (Knowledge), Geography (Knowledge), and History and Archeology (Knowledge).

For some skills, O*NET measures both the level of skill required and the importance of that skill. We use the rating of importance for each skill, which is strongly correlated with level ratings.

One issue in merging O*NET information with ACS data is that O*NET uses 6 digit occupation codes, while some observations in the ACS have occupations listed with only 4 or 5 digits. Because occupation codes are hierarchical, occupations sharing the first 4 or 5 digits have quite

[3]Full descriptions of the variables are available on the O*NET website, onetonline.org.

similar skill requirements, so we impute skill values based on the average among occupations sharing the same non-missing digits. O*NET also does not contain skill measures for military occupations, so we drop such observations.

# 3 Econometric methods

Our research design exploits variation in the timing of accountability reforms. While accountability reforms could plausibly be caused by perceptions about gradual trends in the general quality of education, it is highly unlikely that accountability reforms were targeted in response to perceptions about the quality of specific cohorts of students – e.g., it is improbable that lawmakers would have decided to have accountability in 9th but not 10th grade on the basis of any knowledge of whether rising 10th graders were likely to have better labor market outcomes than rising 9th graders. For this reason, cohorts who just barely missed exposure to accountability are likely to be comparable to cohorts who were just barely exposed.

We implement four different specifications based on this variation. In each regression, an observation is a state-cohort, where state $s$ refers to the birth state and cohort $c$ is defined using the simple assignment rule described in Section 2. We construct state-cohort observations from individual observations by taking the average of each outcome variable (or, for income, the log of the average)[4] among all individuals from that state-cohort in each survey year, then taking the average of these averages across survey years.[5] For each cohort, we use only survey years in which every individual in the cohort would be at least 25 years old.

Our baseline regression is

$$Y_{sc} = \beta\ YrsAcc_{sc} + X'_{sc}\theta + \alpha_s + \gamma_c + \epsilon_{sc}, \tag{1}$$

where $s$ denotes a state of birth and $c$ denotes a school cohort (e.g., students who are estimated to begin first grade in 1992 using the simple assignment rule). $Y_{sc}$ is the average outcome of interest, e.g. the log of average income among people in a state-cohort, $YrsAcc_{sc}$ is the simple measure of exposure to accountability, $X$ is a vector of controls, $\alpha$ and $\gamma$ are fixed effects, and $\epsilon$ is a mean-zero error term.

Regression models of this kind are typically said to make an assumption of "common trends," which in this case would mean that, if they had not had accountability, states with accountability would have experienced the same evolution in outcomes across time as experienced by states which in reality did not have accountability.

However, a recent strand of research highlights that fixed effects models such as ours do not necessarily impose exactly this assumption (e.g. de Chaisemartin and D'Haultfœuille 2020,

[4]Our results are unaffected by using the average of log of income instead, with cutoff points of either zero (such that the log is defined) or modest positive incomes.

[5]While survey years are obviously not directly comparable, this procedure mixes the same survey years for members of the same cohort in any state, and therefore, conditional on the cohort fixed effects in our regression, does not introduce any error correlated with exposure to accountability.

Goodman-Bacon 2021). This is because time trends are identified not only from changes in outcomes among observations which remain untreated between periods, but also from changes in outcomes among observations which remain (equally) treated between periods. This distinction matters if there are time-varying effects of accountability, such that the trend in outcomes if not treated is not necessarily equal to the trend in outcomes conditional on a fixed but non-zero amount of treatment.

To illustrate this in our context, imagine that accountability increases math skills, and that this effect increases over time after implementation. Then early-implementing states will have steep upwards trends in math skills, even once exposure to accountability stabilizes. This biases the time trend estimates upwards, such that the model overestimates the counterfactual level of math skills in late-implementing states. This would bias our estimate of the treatment effect downwards.

The key conditions under which estimates are biased are when (i) there is a staggered implementation of treatment, and (ii) treatment has effects which are heterogeneous over time. Condition (i) is clearly satisfied in our context. Furthermore, condition (ii) is likely to be satisfied as well, since there are many reasons why the effects of accountability might change over time: educators might gradually adapt to new regulations, sanctions often take time to kick in, and, to the extent that accountability programs produce changes in personnel, these changes occur gradually over time. In fact, some previous research suggests that effects on test score performance are to some extent lagged (Linn 2000).

We use two approaches to limit this problem. The first is to restrict our sample to cohorts in which a substantial number of states have not yet implemented consequential accountability. This is accomplished by ending the sample with the cohort which the simple assignment rule predicts would attend first grade in the 1991 school year.[6][7] Restricting the sample in this way ensures that time trends are predominantly estimated using states which have not yet implemented accountability. In Appendix C.1, we demonstrate that this still results in a negative weight on treatment effects for some treated observations, but with relatively small negative weights – though negative weights become more common and larger as control variables are added.

Our second approach is to implement a multi-step procedure that estimates time trends only using states which have not yet experienced consequential accountability. This approach directly imposes the common trends assumption as described above: that, if they had not implemented accountability, trends among states which have experienced accountability would have mirrored trends in states which actually did not.

The multi-step procedure is as follows. In the first step, we estimate the equation

$$Y_{sc} = X'_{sc}\tilde{\theta} + \tilde{\alpha}_s + \tilde{\gamma}_c + \tilde{\epsilon}_{sc}, \tag{2}$$

[6]NCLB results in a dramatic increase in accountability in the following cohorts.

[7]We begin our sample with the 1978 cohort, which is prior to the start of accountability in all states, in order to estimate state fixed effects using relatively recent cohorts prior to accountability, while also ensuring that there exist pre-accountability cohorts for every state.

restricting the regression to only those observations which have not yet experienced accountability (including all observations from states which are never treated during our sample period). The best-fit population parameters might be different from the estimands of baseline regression, hence we use tildes to denote the modified versions. Next, we construct

$$\widetilde{Y}_{sc} := Y_{sc} - (X'_{sc}\hat{\tilde{\theta}} + \hat{\tilde{\alpha}}_s + \hat{\tilde{\gamma}}_c),$$

where hats denote estimates from the first step. Finally, we regress this modified outcome variable on exposure to accountability:

$$\widetilde{Y}_{sc} = \tilde{\beta}\ YrsAcc_{sc} + \tilde{\epsilon}_{sc}. \tag{3}$$

The parameter $\tilde{\beta}$ is interpreted as the effect of exposure to accountability. We estimate standard errors with the wild cluster bootstrap (Cameron et al. 2008, Djogbenou et al. 2019).[8]

A brief word may help explain this procedure. We are interested in asking whether greater quantities of accountability result in greater effects on outcome $Y$. That is, letting $Y_{sc}(a)$ denote the outcome that would prevail in state $s$ and cohort $c$ if $YrsAcc$ were set to $a$, we are interested in regressing the causal effect of accountability, which is $Y_{sc}(YrsAcc_{sc}) - Y_{sc}(0) = Y_{sc} - Y_{sc}(0)$, on $YrsAcc_{sc}$.

This is not feasible because, when $YrsAcc_{sc} \neq 0$, we do not observe $Y_{sc}(0)$. Instead, we estimate $Y_{sc}(0)$ using the first-step regression in Equation 2, such that our eventual regression from Equation 3 above is of $Y_{sc} - \widehat{Y_{sc}(0)} = \widetilde{Y}_{sc}$ on $YrsAcc_{sc}$, with the hat denoting an estimate. The key assumption is therefore that Equation 2 describes $Y(0)$ both for treated and untreated states, i.e. that trends in untreated states reflect what trends would have been in treated states had they been untreated. This assumption matches the qualitative description most commonly given for common trends assumptions.

Our third and fourth specifications additionally account for the fact that the simple measure of accountability is likely inaccurate.

In our third specification, we replace the simple measure of accountability in the baseline regression with the detailed measure of accountability:

$$Y_{sc} = \xi\ ExpAcc_{sc} + X'_{sc}\omega + \psi_s + \phi_c + \nu_{sc}, \tag{4}$$

where $ExpAcc_{sc}$ is the detailed accountability measure for cohort $c$ born in state $s$, $\psi$ and $\phi$ are state and cohort fixed effects, and $\nu$ is a mean-zero error term.

Finally, in our fourth specification, we use the detailed accountability measure while estimating fixed effects and controls only using those cohorts which are not yet assigned to accountability by the simple measure. That is, we estimate

$$\widetilde{Y}_{sc} = \tilde{\xi}\ ExpAcc_{sc} + \tilde{\nu}_{sc}. \tag{5}$$

[8] We implement the wild cluster bootstrap using a Rademacher distribution (Davidson and Flachaire 2008), defining states as the clustered groups. The wild cluster bootstrap gives standard errors very close to the usual clustered standard errors for the more standard version of our panel analysis (first and third specifications).

While technically *ExpAcc* is not exactly equal to 0 for some observations used to construct $\widetilde{Y}_{sc}$, the expected exposure to accountability is so low (see Figure 1) that this is unlikely to be a significant source of bias.

Standard errors for these third and fourth specifications must account both for sampling error in the regression equations (imagining states and accountability laws as being drawn from a superpopulation) and sampling error in the construction of the variable *ExpAcc*. We therefore account for this with a bootstrapping procedure which resamples at both of these stages. In each iteration of the bootstrap, we resample each source of data used to construct *ExpAcc*, then use a wild cluster bootstrap in the "second stage," i.e. in the estimation of Equations 4 or 5.

**Controls** While accountability is unlikely to be related to students' unobservables over short time horizons, accountability might plausibly be related to gradual changes in a state, in which case the common trends assumption might not hold. A key threat is that changes to accountability might be related to demographic changes. We therefore implement some specifications in which we control for the characteristics of parents of each state-cohort. The list of parental controls used includes log of average income, educational attainment, and occupational use of math, writing, critical thinking, creativity, and non-tested skills, for both parents native to the state and parents who migrated there (see below).

Our measures of parental characteristics come from 1990 Census data. This data reports whether the (probable) mother or father lives in the same home as the child; if so, the parent's person number allows for a probabilistic match between parent and child.[9] To limit any attenuation bias which might arise from sampling error in the 1990 Census data (especially among smaller states), we estimate parental characteristics for each state-cohort, then fit a quadratic trend across cohorts within each state, and use the fitted value from this quadratic as our estimate of parental characteristics for that state-cohort. (Leave-one-out testing shows that this delivers a slightly more accurate predictive model of parental characteristics than taking a simple average within the state-cohort, due to the limited number of observations in each state-cohort cell.) Because trends in parents' demographics can represent either migration or changes in the quality of the education system or other inputs which was available to parents – and these two mechanisms might produce different predictions about their children's outcomes – we separately construct trends in parental characteristics for parents who were born in the state in question and for parents who were not born in that state but whose children live there now, and control for these two variables separately.

[9]The probable parent's person number is found via the variables MOMLOC and POPLOC provided by IPUMS. Likely parent-child relationships within households are imputed through the variable RELATE, which lists a respondent's relationship to the householder. See descriptions for variables MOMRULE and/or POPRULE at usa.ipums.org for more detail, including how ambiguous cases where multiple potential parents exist are treated.

# 4 Results

The results for our four specifications are shown in Tables 1 and 2 for education and labor market outcomes, respectively. Each cell reports the coefficient on the accountability measure for a given outcome, specification, and possible inclusion of controls. Rows without controls include only state and cohort fixed effects. Coefficients with a subset of our controls do not depart substantially from the coefficients shown in Tables 1 and 2.

We find that each year of exposure to accountability increases educational attainment by between .01 and .03 years. Effects on high school completion are robust to choice of specification, while effects on college completion are more sensitive. More detailed results by year of attainment are given in Appendix C.2.

The first three panels of Figure 2 show the pattern of educational attainment over time. The horizontal axis is the number of cohorts after the onset of consequential accountability, as in Figure 1, restricting to those states which have positive years of accountability under the simple assignment rule at any point in the sample. The vertical axis is the average outcome net of controls

$$Y_{sc} - X'_{sc}\hat{\theta} - \hat{\alpha}_s - \hat{\gamma}_c,$$

i.e. it is an estimate of the sum of the residual and treatment effect, $\beta\, YrsAcc_{sc} + \epsilon_{sc}$, from our first specification, implemented with the full set of controls. The graphs also show linear fits of the residualized outcome against years since accountability, separately estimated for state-cohorts with negative and non-negative years since accountability; note that this is not precisely the same as the variation used to produce our results in Tables 1 and 2, which additionally exploit variation in accountability among observations with the same number of years since accountability.

The graphs show no pattern of divergence from common trends as cohorts approach the start of accountability policies. This suggests that our results are not likely substantially biased by gradual changes taking place in states implementing accountability. Then there is a roughly linear increase in educational attainment following the implementation of consequential accountability, mirroring the pattern of increasing exposure to accountability from Section 2.

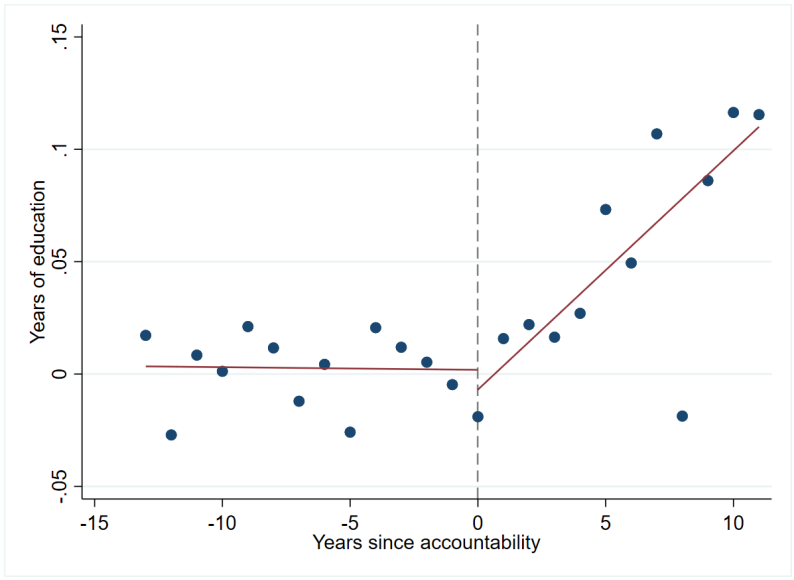
Using log of income as an outcome, we find increases in income of .2-.5% per year of assignment to, or exposure to, accountability. These estimates are not statistically different from 0, except when using controls in Specifications 2 and 4.

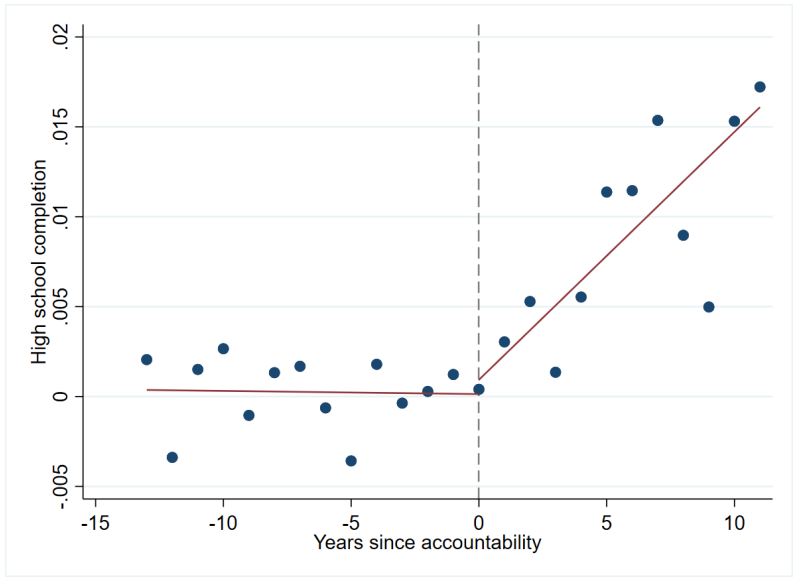
Evidence for effects on the skills most closely measured by standardized tests is mixed. Effects on use of math skills are estimated to be virtually exactly 0. Estimates for writing skills are consistently positive across specifications, with magnitudes on the order of .001-.005 standard deviations per year of accountability. Some but not all of these results are statistically significant.

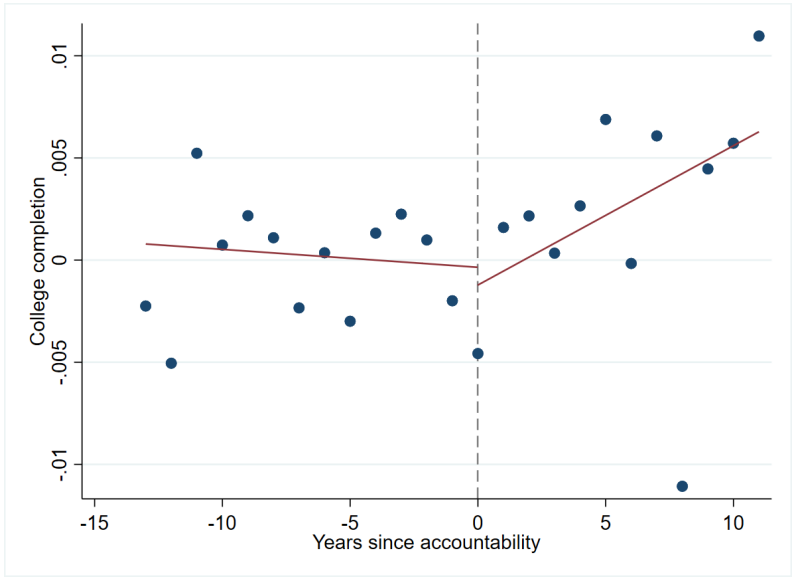
Creativity and critical thinking – measures of broader cognitive function which could in principle be diminished by teaching to the test, or increased by higher-quality instruction – produce estimates which are typically positive and insignificant, and also on the order of thousandths of a standard deviation per year of exposure.

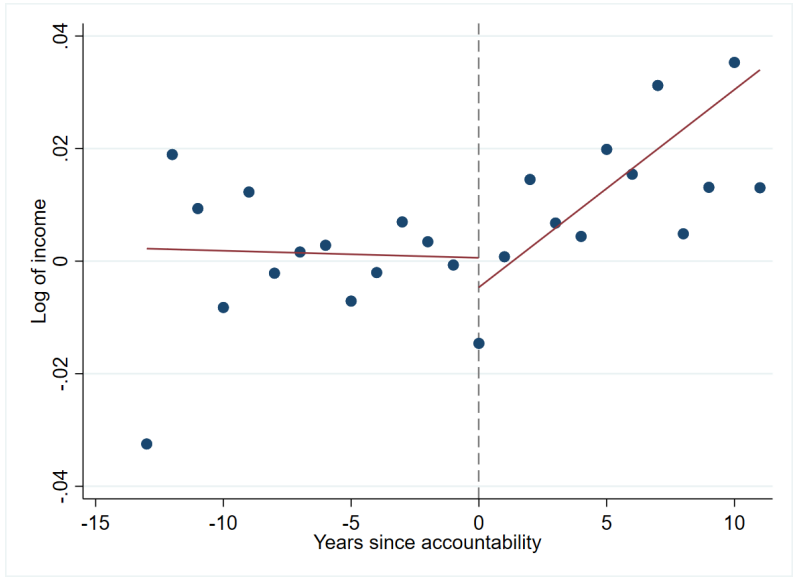
Lastly, skills which might be crowded out by teaching to the test – science and non-tested subjects – show some evidence of negative effects, though with some positive point estimates. However, again, the magnitude of effects can be bounded to a fraction of a hundredth of a standard deviation, and the estimates are almost all insignificant.

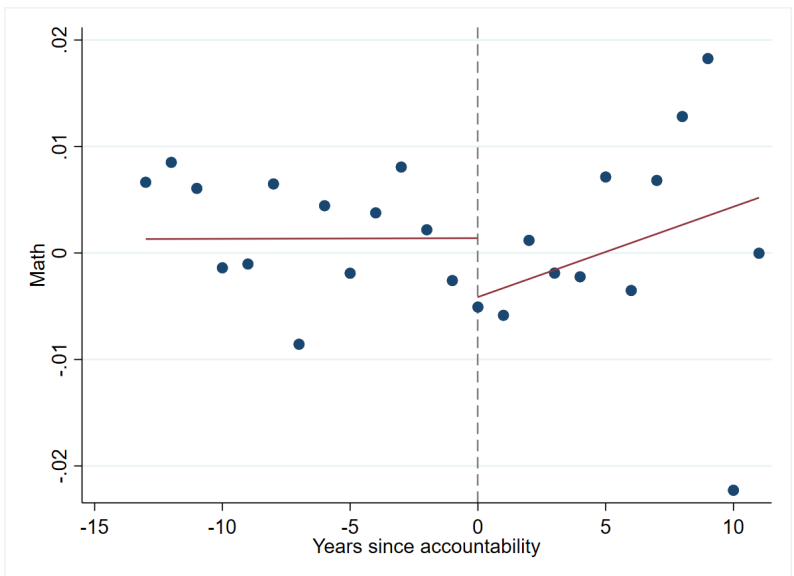
Figure 2: Residualized outcomes by cohort

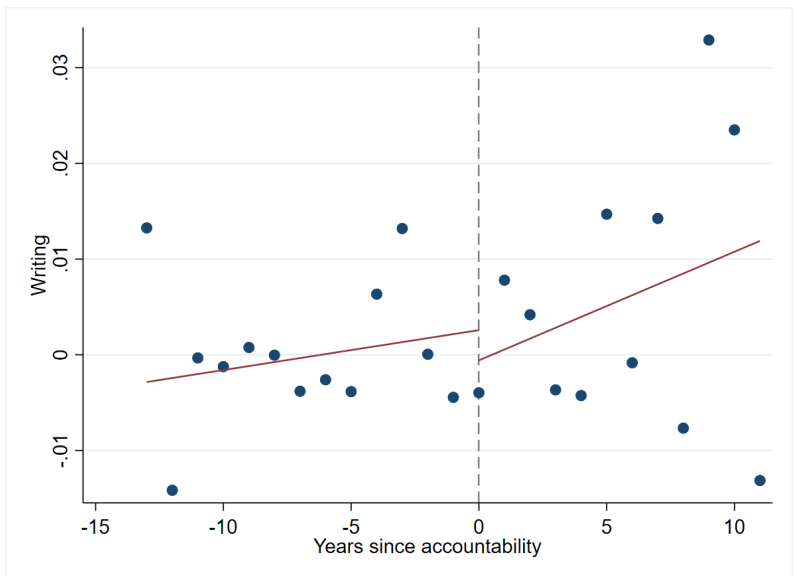


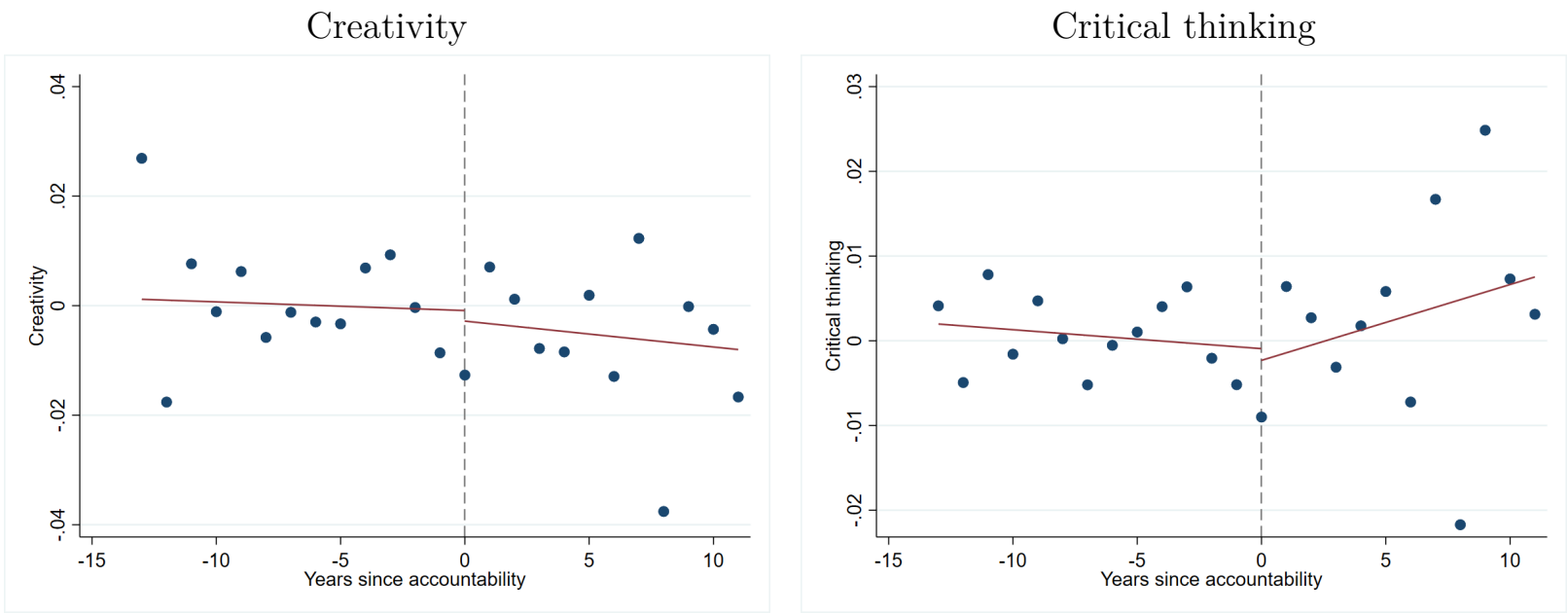






Average outcomes, net of controls, by years since accountability. Averages are across state-cohorts. Outcomes net of controls are outcome minus estimated fixed effects and parental controls (log of average income, average educational attainment, and average occupational skills) controlling separately for average values of parents native to the state and parents who migrated.

Table 1: Effects of accountability on educational attainment

| | (1) Years Educ | (2) Complete HS | (3) Complete BA |
|---|---|---|---|
| **Specification 1:** Standard FE, using simple assignment rule | | | |
| $YrsAcc$ | 0.0195 | 0.0020 | 0.0020 |
| | (0.0046) | (0.0005) | (0.0008) |
| (with controls) | 0.0100 | 0.0019 | 0.0003 |
| | (0.0039) | (0.0005) | (0.0007) |
| **Specification 2:** Multi-step, using simple assignment rule | | | |
| $YrsAcc$ | 0.0203 | 0.0019 | 0.0019 |
| | (.0050) | (0.0004) | (0.0007) |
| (with controls) | 0.0193 | 0.0022 | 0.0017 |
| | (0.0039) | (0.0004) | (0.0007) |
| **Specification 3:** Standard FE, using detailed assignment rule | | | |
| $ExpYears$ | 0.0283 | 0.0029 | 0.0029 |
| | (0.0067) | (0.0006) | (0.0011) |
| (with controls) | 0.0157 | 0.0028 | 0.0006 |
| | (0.0062) | (0.0007) | (0.0011) |
| **Specification 4:** Multi-step, using detailed assignment rule | | | |
| $ExpYears$ | 0.0267 | 0.0026 | 0.0025 |
| | (0.0071) | (0.0006) | (0.0010) |
| (with controls) | 0.0269 | 0.0031 | 0.0024 |
| | (0.0055) | (0.0006) | (0.0010) |
| $N$ | 630 | 630 | 630 |

Each entry is a coefficient estimate from a separate regression, on $YrsAcc$ ("simple" measure) for Specification 1 and 2 and $ExpYears$ ("detailed" measure) in Specifications 3 and 4. All estimates control for state and cohort fixed effects. Rows labeled "with controls" are estimates additionally including parental controls (log of average income, average educational attainment, and average occupational skills) controlling separately for average values of parents native to the state and parents who migrated. See text for details of estimation of Specifications 2 and 4. Standard errors in parentheses. Standard errors in Specifications 1 and 3 are clustered by state, and wild cluster bootstrapped by state in Specifications 2 and 4.

## 4.1 Robustness

We next consider a few potential concerns.

**Additional channels** We would like to interpret our result as the direct effect of accountability on affected students. However, there might be indirect ways that accountability could affect outcomes – for example, if families become more or less likely to place their children in private school, or move to other communities, in response to knowing they were assigned to account-

Table 2: Effects of accountability on labor market outcomes

| | (1) ln(Income) | (2) Writing | (3) Math | (4) Creativity | (5) Thinking | (6) Science | (7) Non-tested |
|---|---|---|---|---|---|---|---|
| **Specification 1:** Standard FE, using simple assignment rule | | | | | | | |
| $YrsAcc$ | 0.0025 | 0.0031 | 0.0008 | 0.0021 | 0.0025 | -0.0004 | 0.0019 |
| | (0.0019) | (0.0018) | (0.0015) | (0.0016) | (0.0017) | (0.0013) | (0.0015) |
| (with controls) | 0.0033 | 0.0012 | 0.0004 | -0.0015 | 0.0004 | -0.0007 | -0.0027 |
| | (0.0023) | (0.0015) | (0.0014) | (0.0012) | (0.0014) | (0.0014) | (0.0013) |
| **Specification 2:** Multi-step, using simple assignment rule | | | | | | | |
| $YrsAcc$ | 0.0024 | 0.0028 | 0.0001 | 0.0019 | 0.0021 | -0.0023 | 0.0017 |
| | (0.0019) | (0.0018 | (0.0014) | (0.0019) | (0.0018) | (0.0017) | (0.0016) |
| (with controls) | 0.0035 | 0.0033 | 0.0001 | 0.0019 | 0.0027 | -0.0013 | 0.0016 |
| | (0.0017) | (0.0015) | (0.0013) | (0.0016) | (0.0015) | (0.0021) | (0.0015) |
| **Specification 3:** Standard FE, using detailed assignment rule | | | | | | | |
| $ExpYears$ | 0.0039 | 0.0046 | 0.0012 | 0.0032 | 0.0037 | -0.0003 | 0.0031 |
| | (0.0026) | (0.0025) | (0.0023) | (0.0024) | (0.0025) | (0.0019) | (0.0019) |
| (with controls) | 0.0051 | 0.0017 | 0.0004 | -0.0019 | 0.0008 | -0.0007 | -0.0033 |
| | (0.0032) | (0.0023) | (0.0021) | (0.0019) | (0.0022) | (0.0020) | (0.0020) |
| **Specification 4:** Multi-step, using detailed assignment rule | | | | | | | |
| $ExpYears$ | 0.0030 | 0.0036 | -0.0000 | 0.0025 | 0.0026 | -0.0030 | 0.0022 |
| | (0.0026) | (0.0025) | (0.0020) | (0.0026) | (0.0025) | (0.0023) | (0.0021) |
| (with controls) | 0.0047 | 0.0046 | -0.0001 | 0.0027 | 0.0039 | -0.0016 | 0.0024 |
| | (0.0023) | (0.0020) | (0.0017) | (0.0022) | (0.0021) | (0.0027) | (0.0020) |
| $N$ | 630 | 630 | 630 | 630 | 630 | 630 | 630 |

Each entry is a coefficient estimate from a separate regression, on $YrsAcc$ ("simple" measure) for Specification 1 and 2 and $ExpYears$ ("detailed" measure) in Specifications 3 and 4. All estimates control for state and cohort fixed effects. Rows labeled "with controls" are estimates additionally including parental controls (log of average income, average educational attainment, and average occupational skills) controlling separately for average values of parents native to the state and parents who migrated. See text for details of estimation of Specifications 2 and 4. Standard errors in parentheses. Standard errors in Specifications 1 and 3 are clustered by state, and wild cluster bootstrapped by state in Specifications 2 and 4.

ability. These mechanisms would make it such that even students who would never experience accountability, regardless of whether they were assigned to it or not, could still experience some change in outcomes from being assigned to accountability. In Appendix C.3, we look for evidence of such effects. We find no statistically significant effect on private school attendance. We do estimate that assignment to accountability decreased the probability that students lived in their birth state, but this effect is so small (smaller than .002 change in the probability of living in the birth state) that it could not appreciably affect our results without the effects of such moves on our outcome variables being implausibly large.

**Sample selection** A related concern is that inclusion in the sample might be affected by treatment. Not everyone participates in the labor force, and our results would be biased by sample selection if accountability affects the probability that an individual has an occupation. (Our construction of the income variable permits true zeroes, but our results are not sensitive to dropping observations with zero or very low income.) We investigate this in Appendix C.4 and estimate that accountability had a precise zero effect on the probability of working and of having an observed occupation.

**Robustness to exclusion of control observations** We construct our measures of exposure to accountability to capture the number of years that a student's test score had consequences attached to it. This definition means that the control group potentially contains two groups who plausibly could have been in some ways exposed to accountability.

First, students exposed to accountability programs which were not consequential are labeled as control observations, on the basis of the literature's finding that these programs seem to have produced negligible effects on test scores. However, it is possible that report card accountability programs may still have had an effect on long-run outcomes.

Second, students who were still in school at the time of implementation of a consequential accountability program, but who were not in a school cohort such that test scores had consequences attached, are labeled as control observations. However, it is possible that these students were affected by the implementation of consequential accountability. For example, if consequential accountability is introduced in grades 6-9, the students who are enrolled in 10th grade at that time would be labeled as having no exposure to accountability. Yet, if accountability induced cultural changes, or resulted in the reallocation of resources across grades, or if such students were enrolled in courses with students who were exposed to accountability, then these students might have been affected by the implementation of accountability.

In Appendix C.5, we report results dropping these two categories of observations from the control group. The results are similar to those in Tables 1 and 2, though permutations of these two sample restrictions do produce some estimates either slightly above or below the range of results in Tables 1 and 2.

## 4.2 Interpretation

A natural conceptual question related to our results is whether we should think of the range of values which cannot be statistically rejected as containing large or small effects of accountability. We suggest that the answer probably somewhat depends on the purposes of the estimate.

From the perspective of a cost-benefit analysis, it is likely easy to justify an accountability program which increases students' incomes by .1 or .2%, and certainly by more than that. As an example parameterization, at a 3% discount rate, the present discounted value of increasing earnings by .1% for a student who would otherwise earn exactly $60,000 per year for 40 years starting ten years after leaving the classroom is $1,031; or, for a classroom of 25 such students, $25,799. Effects on the high end of our estimates, like a .5% increase in income per year of exposure, therefore suggest a present discounted value of over $100,000 per classroom per year.

On the other hand, in most specifications, we cannot statistically rule out a zero effect on income. Further, given there is an effect on attainment, we are not in any specification able to rule out the effect on income is equal to what would be expected to arise due simply to the fact that students remain in school longer. If the effect on attainment is .02 years per year of exposure, then, assuming each year of schooling increases earnings by about 5% (Card 1999, Card 2001, Oreopoulos 2006), we would expect to see incomes increase by .1% through this mechanism alone. If all effects on income occurred solely through attainment, it might be just as desirable to simply increase the compulsory schooling age.

More broadly, we might ask whether accountability policies were transformative to children's human capital accumulation, as it was sometimes suggested that they would be. Our estimates suggest not. For comparison to our findings, Chetty and Hendren (2018) estimate that growing up in a one standard deviation better county for producing incomes in adulthood raises income by nearly 10%.[10] Relative to those numbers, our estimates are small: A high-profile education reform involving multiple grades of consequential accountability is only estimated to increase place effects on income by maybe a tenth of a standard deviation. If place effects were normally distributed, for example, this would mean that implementing such a consequential accountability regime would move a community from the 50th to the 54th percentile, or from the 75th to the 78th percentile. Similarly, the effects of accountability are small relative to cohort effects; e.g. IQ has been increasing in the United States by roughly .02 standard deviations per year (Flynn 1984), meaning that the effect of five years of exposure to consequential accountability on cognitive skills is likely smaller in standard deviations than the average year-over-year change in IQ. So, despite being one of the most high-profile education reforms in recent decades, consequential accountability probably explains very little about differences in human capital production across time and place.

[10] Specifically, they estimate that this increases incomes by 10% for children from low-income families, and 6% for children from high-income families.

# 5 Conclusion

We find that consequential accountability programs produced increases in educational attainment of roughly .02 years of attainment per year of accountability. We find some evidence of positive effects on income in adulthood, though not significant in most specifications. Furthermore, effects on the use of skills in adult occupation are generally not significant, and are no larger than a fraction of a hundredth of a standard deviation per year of accountability for each skill measure. Our estimates suggest that accountability programs were probably net beneficial, but they do not appear to have had a transformative effect on human capital production.

# References


[1] Barksdale-Ladd, M., and K. Thomas (2000). WhatâĂŹs at Stake in High-Stakes Testing: Teachers and Parents Speak Out, *Journal of Teacher Education*, 51(5), 384-397.

[2] Cameron, A., Gelbach, J., and D. Miller (2008). Bootstrap-Based Improvements for Inference with Clustered Errors, *The Review of Economics and Statistics*, 90(3), 414-427.

[3] Card, D. (1999). The Causal Effect of Education on Earnings, *Handbook of Labor Economics*, 3(30) 1999, 1801-1863.

[4] Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems, *Econometrica*, 69(5), 1127-1160.

[5] Carnoy, M., and S. Loeb (2002). Does External Accountability Effect Student Outcomes? A Cross-State Analysis, *Educational Evaluation and Policty Analysis*, 24(4), 305-331.

[6] Chetty, R. and N. Hendren (2018). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates, *The Quarterly Journal of Economics*, 133(3), 1163-1228.

[7] Clotfelter, C., Ladd, H., Vigdor, J., and R. Diaz (2004). Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? *Journal of Policy Analysis and Management*, 23(2), 251-271.

[8] Davidson, R., and E. Flachaire (2008). The wild bootstrap, tamed at last, *Journal of Econometrics*, 146, 162-169.

[9] Djogbenou, A., MacKinnon, J., and M. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors, *Journal of Econometrics*, 212(2), 393-412.

[10] de Chaisemartin, C., and X. D'HaultfÅŞuille (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects, *American Economic Review*, 110(9), 2964-2996.

[11] Dee, T., and B. Jacob (2011). The Impact of No Child Left Behind on Student Achievement, *Journal of Policy Analysis and Management*, 30(3), 418-446.

[12] Dee, T., Jacob, B., and N. Schwartz (2013). The Effects of NCLB on School Resources and Practices, *Educational Evaluation and Policy Analysis*, 35(2), 252-279.

[13] Deming, D.J., Cohodes, S., Jennings, J., and C. Jencks (2016). School Accountability, Post-Secondary Attainment and Earnings, *Review of Economics and Statistics*, 98(5), 848-862.

[14] Feng, L., Figlio, D., and T. Sass (2018). School accountability and teacher mobility, *Journal of Urban Economics*, 103, 1-17.

[15] Flynn, J. (1984). The mean IQ of Americans: Massive gains 1932 to 1978, *Psychological Bulletin*, 95(1), 29-51.

[16] Figlio, David, and S. Loeb (2011). School Accountability, *Handbook of the Economics of Education*, 3(8), 383-421.

[17] Figlio, D., and M. Lucas (2004). What's in a Grade? School Report Cards and the Housing Market, *The American Economic Review*, 94(3), 591-604.

[18] Figlio, D., and J. Winicki (2005). Food for thought: the effects of school accountability plans on school nutrition, *Journal of Public Economics*, 89(2-3), 381-394.

[19] Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing, *Journal of Econometrics*, forthcoming.

[20] Greene, J. (2001). An Evaluation of the Florida A-Plus Accountability and School Choice Program, *Manhattan Institute for Policy Research.*

[21] Hanushek, E., and M. Raymond (2001). The Confusing World of Educational Accountability, *National Tax Journal*, 54(2), 365-384.

[22] Hanushek, E., and M. Raymond (2005). Does school accountability lead to improved student performance?, *Journal of Policy Analysis and Management*, 24(2), 297-327.

[23] Hoffman, J., Assaf, L., and S. Paris (2001). High-Stakes Testing in Reading: Today in Texas, Tomorrow? *The Reading Teacher*, 54(5), 482-492.

[24] Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools, Journal of Public Economics, 89(5-6), 761-796.

[25] Jacob, B., and S. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating, *The Quarterly Journal of Economics*, 118(3), 843-877.

[26] Koretz, D., and S. Barron (1998), The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS), *RAND Corporation, MR-1014-EDU.*

[27] Ladd, H. (1999). The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes, *Economics of Education Review*, 18(1), 1-16.

[28] Lavy, V. (2020). Teachers' Pay for Performance in the Long-Run: The Dynamic Pattern of Treatment Effects on Students' Educational and Labor Market Outcomes in Adulthood, *Review of Economic Studies*, 87(5), 2322-2355.

[29] Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies, *Review of Educational Research*, 78(3), 608-644.

[30] Linn, R. (2000) Assessments and Accountability, *Educational Researcher*, 29(2), 4-16.

[31] Loeb, S., and Cunha, J. (2007). Have assessment-based accountability reforms influenced the career decisions of teachers?, *A report commissioned by the U.S. Congress as part of Title I, Part E, Section 1503 of the No Child Left Behind Act of 2001.*

[32] McKenzie, W., and Kress, S. (2015). The Big Idea of School Accountability, *George W. Bush Institute.*

[33] National Research Council (2011). Incentives and Test-Based Accountability in Education, *The National Academies Press.*

[34] Oreopoulos, P. (2006). Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter, *The American Economic Review*, 96(1), 152-175.

[35] Richards, C., and T. Sheu (1992). The South Carolina school incentive reward program: A policy analysis, *Economics of Education Review*, 11(1), 71-86.

[36] Roy, A. (1951). Some Thoughts on the Distribution of Earnings, *Oxford Economic Papers*, 3(2), 135-146.

[37] Shepard, L.A., and K. Dougherty (1991). Effects of high-stakes testing on instruction, *Paper presented at the annual meeting of the American Educational Research Association*, April.

[38] Stecher, B., Barron, S., Kaganoff, T., and J. Goodwin (1998). The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing, *Center for Research on Evaluation, Standards and Student Testing*, CSE Technical Report 482.

[39] Wong, K. (2008). Looking Beyond Test Score Gains: State AccountabilityÕs Effect on Educational Attainment and Labor Market Outcomes, working paper.

[40] Wong, M., Cook, T., and P. Steiner (2015). Adding Design Elements to Improve Time Series Designs: No Child Left Behind as an Example of Causal Pattern-Matching, *Journal of Research on Educational Effectiveness*, 8(2), 245-279.

# Appendix

# A Accountability implementation by state

The following is a description of accountability laws in each state.

## Alabama

In 1995, Alabama passed legislation (Alabama Act 1995-313) for a State School Accountability Plan. The first accountability ratings were given in June 1996. They implemented a three-year cycle of accountability using norm-referenced testing in grades 3-11, among other indicators. After No Child Left Behind was passed, they changed their assessment system and tested grades 3-8 and 11 starting in the 2004 academic year.

## Alaska

Alaska used norm-referenced testing prior to NCLB in grades 3-10 and passed a 1998 law (the Quality Schools Initiative) mandating the development and implementation of an accountability system. However, the legislature later delayed implementation of key accountability measures until after the start of NCLB (HB 352 of the 22nd legislature).

## Arizona

In November 2001, Arizona passed legislation (Proposition 301, section 15-241 of the Arizona Revised Statutes) which established a system of student assessment (AZ LEARNS). Testing was implemented in grades 3, 5, 8, and 10. Arizona made grade changes to comply with NCLB standards in 2006.

## Arkansas

Arkansas passed ACT 999 of 1999 which implemented a consequential accountability program starting in 2000. Initially, grades 4, 6, 8, and 11 were tested, with subsequent changes coming after the passage of NCLB. Arkansas used End-of-Course tests in accountability, so the most logical grade was assigned to each test used.

## California

California passed the Public Schools Accountability Act of 1999, which implemented consequential accountability starting in 2000.

## Colorado

We rate NCLB as the first exposure to accountability in Colorado. Prior to NCLB, Colorado implemented their Colorado Student Assessment program, established by House Bill 93-1313. By 2001, grades 3-10 were being tested in some capacity. However, this assessment system did not have any stakes attached outside of report cards for schools, so we do not rate this accountability system as consequential.

## Connecticut

Connecticut implemented the Connecticut Mastery Tests in 1985 to identify weak schools. However, there was no formal accountability attached to these tests. We therefore label Connecticut as starting accountability with NCLB.

## Delaware

Delaware passed the Accountability Act of 1998 to design a formal assessment system and accountability. However, the law provided only weak incentives attached to testing performance, such as recognition for high-performing schools. Thus, we treat NCLB as the first implementation of consequential accountability for Delaware.

## Florida

Florida implemented the FCAT testing system for the first time in 1998 for grades 4, 5 , 8, and 10. School-level consequences were attached to performance in the following year via the implementation of Florida's 1999 A+ Plan for Education. Grades were subsequently added in 2001 (3-10).

## Georgia

Georgia passed the A Plus Education Reform Act in 2000 but did not have consequential accountability actually implemented prior to NCLB.

## Hawaii

Hawaii did not have consequential accountability prior to NCLB, though it did publish test scores starting in 2001.

## Idaho

Idaho did not have consequential accountability prior to NCLB.

## Illinois

Illinois had an accountability system for a number of years prior to NCLB, but we judge that the stakes and enforcement were insufficient to be labeled as consequential. The Chicago Public School System implemented a consequential accountability system in 1996. However, since this system was not statewide, Illinois is treated as having no state-wide consequential accountability exposure prior to NCLB.

## Indiana

Indiana passed legislation for a performanced-based accountability program in 1999 (Public Law 221-1999). However, the program wasn't fully implemented until the 2005 academic year. Thus, NCLB was the first performance-based accountability in Indiana.

## Iowa

Iowa did not have an accountability program prior to NCLB.

## Kansas

Kansas did not have consequential accountability prior to NCLB, though they did have report cards starting in 1995.

## Kentucky

Kentucky passed the Kentucky Education Reform Act of 1990 to implement their assessment and accountability system. Rewards and sanctions were conditional on the outcome of three-year cycles of testing, the first of which started in the 1992 school year. Kentucky made changes to its assessment program (KIRIS) over time and in response to NCLB.

## Louisiana

Louisiana implemented accountability with their new assessment system LEAP 21 in 1999 to accompany new content standards established by the state. Grades assessed were expanded in a rollout over several years.

## Maine

Maine had report card accountability by 1999 but no consequences were attached to their testing.

## Maryland

Maryland had the Maryland School Performance Program since 1989 and began administering awards based on 1996 test results on statewide assessments in grades 3, 5, 8, 9, and 11. Schools were also eligible for reconstitution as well. When established, the High School Assessment testing was treated as accountability testing in grade 10 given the subjects covered.

## Massachusetts

Massachusetts implemented accountability following the Massachusetts Education Reform Act of 1993. The Massachusetts Comprehensive Assessment System (MCAS) started testing in 1998 in grades 4, 8 and 10, with more grades added in 2001 and in response to NCLB.

## Michigan

Michigan first implemented an accreditation program that made schools subject to potential sanctions for assessment scores (MEAP) in 1995 (amendment to Public Act 25 of 1990). However, this system was quite weak and was subsequently overhauled in time to apply to the 1999 MEAP results, with the state having takeover power and the ability to close schools.

## Minnesota

Minnesota had a report card system established in 1996 but without any consequences attached prior to NCLB.

## Mississippi

Mississippi had an accreditation system established by the Education Reform Act of 1982. Subsequent legislation in 1994 attempted to bolster the system, but it wasn't until the Mississippi Student Achievement Improvement Act of 1999 that a more typical consequential accountability system was proposed. This program was not in place until 2003. Prior to this there was some district-level accountability, though with consequences limited to schools providing a plan for improvement. Based on this, Mississippi was not counted as having consequential accountability prior to NCLB. Mississippi used End-of-Course testing at the high school level; based on the specific courses tested, we label its post-NCLB accountability as occurring in grades 9 and 10.

## Missouri

Missouri did not have a consequential accountability program prior to NLCB.

## Montana

Montana did not have a consequential accountability program prior to NLCB.

## Nevada

Nevada had some accountability laws as far back as 1989, and published school report cards starting in 1995. The Nevada Education Reform Act of 1997 added some intervention and possibility of replacing administration at the school level tied to student achievement. The 1998 school year was the first year of consequential accountability with initial grades tested of 4, 8, 10, and 11.

## New Hampshire

New Hampshire did not have a consequential accountability program prior to NCLB.

## New Jersey

New Jersey was dropped from the main analysis due to the decentralized nature of its school accountability. Starting in 1987, the state had an ability to take over low-performing school districts with testing having been part of the determination. In 1995, New Jersey started publishing school report cards and a year later passed the Comprehensive Educational Improvement and Financing Act of 1996 which established some statewide assessment standards for students. However, much of the standards, curriculum and pupil funding were still defined at the district level, with large discrepancies across districts. Districts falling behind on standards were dealt with on a case-by-case basis with no uniform sanctions across districts. Because of the lack of a clearly defined statewide system of consequences, we do not view this as a statewide consequential accountability system. In our detailed assignment variable, we treat New Jersey as having no consequential accountability prior to 2003.

## New Mexico

New Mexico was dropped from the main analysis due to the ambiguity of whether the state's accountability program was consequential prior to NCLB. Starting with the passage of the Incentives for School Improvement Award program in 1997, New Mexico started having some small monetary incentives for school improvement. However, these awards could not be given in the form of teaching salaries, so the incentives were limited. Also, while there was some threat of intervention, it appears to never have been actually implemented. For purposes of our detailed assignment rule, we treat New Mexico as not having accountability prior to NCLB.

## New York

New York was dropped from the main analysis due to having too gradual a pace of accountability reform. New York had begun their Schools Under Registration Review (SURR) program in 1989, allowing the takeover of low-performing schools. While this did eventually result in the takeover of some schools, this process was not governed by hard and fast rules and many schools on the SURR listings remained there through much of the 1990s. New York introduced report cards in 1996 and implemented more accountability measures in 1999 via their System of Accountability for Student Success, which expanded on SURR. Because of the takeover power granted under SURR, for the purposes of the detailed assignment rule, we designate accountability in New York as starting with the establishment of the SURR program in 1989.

## North Carolina

North Carolina passed the ABCs of Public Education in March of 1995. This resulted in a consequential accountability program which started in the 1997 school year testing grades 3-8 before adding high school testing in grade 10 a year later. A limited set of school districts implemented the accountability system in the 1996 school year, but we date the start of accountability to the statewide implementation the following school year.

## North Dakota

North Dakota did not have a consequential accountability program prior to NCLB.

## Ohio

Ohio established a district-level accountability program via the 1997 legislation House Bill 55. This 1997 reform included report cards but did not provide any rewards or sanctions for schools or districts. Therefore, we label Ohio as having no consequential accountability system prior to NCLB.

## Oklahoma

Oklahoma was dropped from the main analysis due to relatively weak enforcement and lack of clarity in assessment systems. The Oklahoma Educational Indicators Program was established in 1989, and a subsequent 1996 law mandated testing with takeover ability on a three year cycle. However, enforcement appears to have been limited in practice, and other sources (e.g., Carnoy and Loeb 2002) characterize their accountability system as weak. For the purposes of the detailed assignment rule, we label Oklahoma as having consequential accountability starting in 1996.

## Oregon

Oregon established an accountability system in 2001 with report cards published at the school and district level. Additionally, if districts underperformed according to specified achievement and improvement standards, it was within the state's power to withhold state funding, spending authority of federal funds, and allow students to transfer schools and/or districts. Thus, we treat Oregon as having consequential accountability prior to NCLB. Initial testing was in grades 3, 5, 8, and 10.

## Pennsylvania

Pennsylvania was dropped from the main analysis due to ambiguity about the level of incentives for teachers in the state. Pennsylvania established a performance funding program in 1998 based off of results from the Pennsylvania System of School Assessment. Just 25% of the awarded funds were eligible for teacher bonuses, however, which amounted to a relatively weak incentive for teachers. Furthermore, Pennsylvania did not establish other elements of a consequential accountability system until after the passage of NCLB. As a result, Carnoy and Loeb (2002) characterize this as a weak incentive system. Given there were financial incentives prior to NCLB, but those incentives were weak, we judged it to be ambiguous whether Pennsylvania should properly be categorized as having consequential accountability. For purposes of our detailed assignment rule, we treat Pennsylvania as not having consequential accountability prior to NCLB.

## Rhode Island

Rhode Island implemented consequential accountability in the 1998 academic year; see amendments made to the Rhode Island Student Investment Act (Article 31 Sub A). Grades tested initially were 4, 8, and 10.

## South Carolina

South Carolina implemented consequential accountability through the Education Accountability Act for the 1999 school year, testing grades 3-8 and 10.

## South Dakota

South Dakota did not have a consequential accountability program prior to NCLB.

## Tennessee

Tennessee implemented a rewards program and probationary status via the Tennessee Goals and Performance program in 1992. The program in part used a value-added system (TVAAS) with a

3 year cycle for student testing. Monetary rewards could be used int he form of teacher bonuses and probationary status came with the threat of sanctions for repeated poor performance.

## Texas

Texas implemented a consequential accountability program in the 1994 school year with the Texas Assessment of Academic Skills program (TAAS). Texas assessed grades 3-8 and 10 at the start.

## Utah

Utah did not have a consequential accountability program prior to NCLB.

## Vermont

Vermont had report card accountability prior to NCLB but not consequential accountability.

## Virginia

Virginia implemented a performance component to their school accreditation program in 1999 via their Standards of Learning assessments. Repeated low performance resulted in state interventions via improvement plans and potential loss of accreditation.

## Washington

Washington had report card accountability prior to the passage of NCLB but not consequential accountability.

## West Virginia

West Virginia implemented their Performance-based Accreditation Program in 1992. State intervention and possible funding loss was tied to this program. The program was reformed with new testing in 1996.

## Wisconsin

Wisconsin did not have a consequential accountability program prior to NCLB.

## Wyoming

Wisconsin had report card accountability prior to the passage of NCLB but not consequential accountability.

Table A1

| State | Year Started* | Initial Grades Tested | Grade Changes | Included in Main Analysis? |
|---|---|---|---|---|
| Alabama | 1996 | 3-11 | 3-8,11 (2004) | Y |
| Alaska | 2003 | 3-10 | | Y |
| Arizona | 2003 | 3,5,8,10 | 3-8, 10 (2006) | Y |
| Arkansas | 2001 | 4,6,8,11 | 3-8,10 (2006) | Y |
| California | 2000 | 2-11 | | Y |
| Colorado | 2003 | 3-10 | | Y |
| Connecticut | 2003 | 4,6,8,10 | 3-8,10 (2006) | Y |
| Delaware | 2003 | 3-6,8,10-11 | 2-11 (2006) | Y |
| Florida | 1999 | 4,5,8,10 | 3-10 (2001) | Y |
| Georgia | 2003 | 3-8,11 | | Y |
| Hawaii | 2003 | 3,5,8,10 | 3-8,10 (2006) | Y |
| Idaho | 2003 | 4,8,10 | 3,4,7,8,10 (2004); 3-8,10 (2005) | Y |
| Illinois | 2003 | 3-5,7,8,11 | 3-8,11 (2–6) | Y |
| Indiana | 2003 | 3,6,8,10 | 3-10 (2004) | Y |
| Iowa | 2003 | 4,8,11 | 3-8,11 (2006) | Y |
| Kansas | 2003 | 4,5,7,8,10,11 | 3-8,10 (2006) | Y |
| Kentucky | 1992 | 4,7,8,12 | 4-5,7-8,10-12 (1999); 3-8,10-11 (2007) | Y |
| Louisiana | 1999 | 3-8 | 3-10 (2001); 3-11 (2002) | Y |
| Maine | 2003 | 4,8,11 | 3-8,11 (2006) | Y |
| Maryland | 1996 | 3,5,8,9,11 | 3,5,8, HSA(10) (2000); 3-8, HSA(10) (2005) | Y |
| Massachusetts | 1998 | 4,8,10 | 3-4,6-8,10 (2001); 3-8, 10 (2006) | Y |
| Michigan | 1999 | 4,5,7,8,11 | 3-8,11 (2006) | Y |
| Minnesota | 2003 | 3,5,8,10-11 | 3,5,7-8,10-11 (2004); 3-8,10-11 (2006) | Y |
| Mississippi | 2003 | 3-8, EOCs(9,10) | | Y |
| Missouri | 2003 | 3-8,EOCs(9,10) | | Y |

| | | | | |
|---|---|---|---|---|
| Montana | 2003 | 4,8,11 | 4,8,10 (2004); 3-8,11 (2006) | Y |
| Nebraska | 2003 | 4,8,11 | 3-8,11 (2006) | Y |
| Nevada | 1998 | 4,8,10,11 | 2-5,8,10-11 (2002); 3-8,10-11(2006) | Y |
| New Hampshire | 2003 | 3,6,10 | 3-8,10 (2006); 3-8,11 (2008) | Y |
| New Jersey | 2003 | 4,8,11 | 3,4,8,11 (2005); 3-8,11 (2006) | N |
| New Mexico | 2003 | 3-9 | 3-9,11 (2005); 3-8,11 (2008) | N |
| New York | 1989 | 3,6,9 | 4,8,9-12 (1999); 3-12 (2006) | N |
| North Carolina | 1997 | 3-8 | 3-8,10 (1998) | Y |
| North Dakota | 2003 | 4,8,12 | 3,4,5,7,8,10 (2005); 3-8,10 (2006) | Y |
| Ohio | 2003 | 4,6,9,12 | 3,4,5,7,8,10 (2004); 3-8,10 (2005) | Y |
| Oklahoma | 1996 | 3,5,7,8,11 | 3,5,7,8,EOI's -9,10 (2002); 3-5,7,8,EOI's-9,10 (2005); 3-8,EOI's-9,10 (2006) | N |
| Oregon | 2000 | 3,5,8,10 | 3-8,10 (2006) | Y |
| Pennsylvania | 2003 | 5,8,11 | 3,5,8,11 (2005); 3-8,11 (2006) | N |
| Rhode Island | 1998 | 4,8,10 | 3,4,7,8,10 (1999); 3,4,7,8,11 (2004); 3-8,11 (2006) | Y |
| South Carolina | 1999 | 3-8,10 | | Y |
| South Dakota | 2003 | 3-8,11 | | Y |
| Tennessee | 1993 | 3-8 | 3-12 (2001); 3-10 (2004) | Y |
| Texas | 1994 | 3-8,10 | | Y |

| | | | | |
|---|---|---|---|---|
| Utah | 2003 | 3-11 | | Y |
| Vermont | 2003 | 2,4,6,8,10,11 | 3-8 (2006); 3-8,11 (2008) | Y |
| Virginia | 1999 | 3,5,8,EOC's-9-11 | 3-8,EOC's-9-11 (2006) | Y |
| Washington | 2003 | 4,7,10 | 3-8,10 (2006) | Y |
| West Virginia | 1992 | 3,6,9,11 | 3-11 (1996); 3-8,10 (2004) | Y |
| Wisconsin | (2003) | 4,8,10 | 3-8,10 (2006) | Y |
| Wyoming | (2003) | 4,8,11 | 3-8,11 (2006) | Y |

*Years listed represent the end of an academic year, i.e. 1997 indicates the 1996-1997 academic year.

# B Construction of measurement error estimates

In order to construct the detailed measure of exposure to accountability, we must account for three reasons why an individual's exposure to accountability cannot be directly measured from their age and state of birth. First, a student may have moved from their birthplace and attended public school elsewhere. Second, students may not have attended a public school in a given grade, either because they had left school already or because they attended private schools, which are not subject to consequential accountability rules. Third, students may not be members of the school cohort we believe them to be in.

This appendix describes how we construct an expected years of accountability variable by accounting for each of these different sources of measurement error.

We construct the expected number of years of true accountability for individual $i$ as

$$\sum_{gst} D_{gst} * P_{gst}(bstate_i, bquarter_i, byear_i),$$

where $D_{gst}$ is a dummy equal to one if there is accountability in grade $g$ public schools in state $s$ during school year $t$, and $P_{gst}(bquarter_i, byear_i, bstate_i)$ is the probability that an individual with $i$'s quarter, year, and state of birth would attend a public school in grade $g$ in state $s$ during year $t$.

$D_{gst}$ is defined on the basis of accountability laws, as described in Appendix A.

$P_{gst}(bstate_i, bquarter_i, byear_i)$, however, must be estimated. This probability is equal to the product of the probabilities that (i) student $i$ is in grade $g$ at time $t$, i.e., the probability that $i$ is in a particular school cohort; (ii) $i$ lives in state $s$ conditional on being in grade $g$ at time $t$; and (iii) $i$ attends a public school conditional on living in state $s$ and being in grade $g$ at time $t$. For simplicity, we will refer to these as probabilities $P^1$, $P^2$, and $P^3$ respectively.

**Cohort and grade assignment** We use the birth year and birth quarter in ACS data to generate a starting point for our estimates of $P^1$. The birth year variable is a crude calculation of respondent's age subtracted from the year of survey. Assuming a uniform distribution for birthdays and time of survey (ACS surveys take place throughout the year), there is a fifty-percent chance respectively of the birth year calculated to be correct and for it be ahead by one year. Adding to this, an observation may be in a different grade by being held back, moved forward, or starting school earlier or later than expected.

The 1980 5% Census, however, has an exact time of survey (April 1st) which can be combined with age and birth quarter to back out respondents' exact year and quarter of birth. Furthermore, the 1980 Census reports the grades that students are in. This allows us to estimate the probability that a student of a given age and quarter of birth would be in a given school cohort, i.e., we can estimate the probability that a student is in the cohort that we assumed for the simple assignment rule, or is 1 year behind, 2 years behind, etc. We estimate probabilities of being a fixed number of years ahead of or behind the assumed school cohort separately for early (K-4), middle (5-8), and high school (9-12) grades, since we find little evidence that these probabilities vary significantly over local ranges of grades. We also do not find evidence that these probabilities differ appreciably by state of residence or state of birth, so we assume a single probability for each cluster of grades.

Next, we adjust this for the probability that the respondent's birth year is mismeasured, which is a function of quarter of birth. For example, assuming interviews are distributed evenly across the year, there is a 7/8 chance that a randomly selected respondent born in the fourth quarter is surveyed before their birthday; while, among respondents born in the first quarter, there is only a 1/8 chance that the interview is conducted before their birthday.

**Migration patterns** For simplicity, we assume that $P^2$ does not depend on $t$, but that there is instead a constant probability that individuals born in state $s$ would live in state $s'$ during grade $g$.

The 1990 and 2000 5% Census samples report state of birth and current state. While these samples do not have a measure of the exact grade a student is in, we can use year and quarter of birth to estimate a most likely grade. From this information, we can estimate the probability that someone born in state $s$ lives in state $s'$ in grade $g$ for any $s$, $s'$, and $g$. To reduce the role of sampling error in these estimates, we group grades into early (K-4), middle (5-8), and high school (9-12) levels – e.g., we assume that the probability that someone born in North Carolina lives in Montana is the same in 5th grade as in 8th grade.

**Attending public school** For simplicity, we assume that $P^3$ does not depend on $t$. However, we do allow it to vary by state of birth, state of current residence, and grade.

The 1990 and 2000 5% Censuses report whether the respondent attends a private or public school (or is not in school). We estimate the probability of attending a public school conditional on state of birth, current state, and grade using a logit model with fixed effects for each of these

three variables. Because the data do not contain exact grade, we use estimated grade based on age and quarter of birth. The fitted values from the logit are then used as estimated values of $P^3$.

# C Additional results

## C.1 Weights on treatment effects

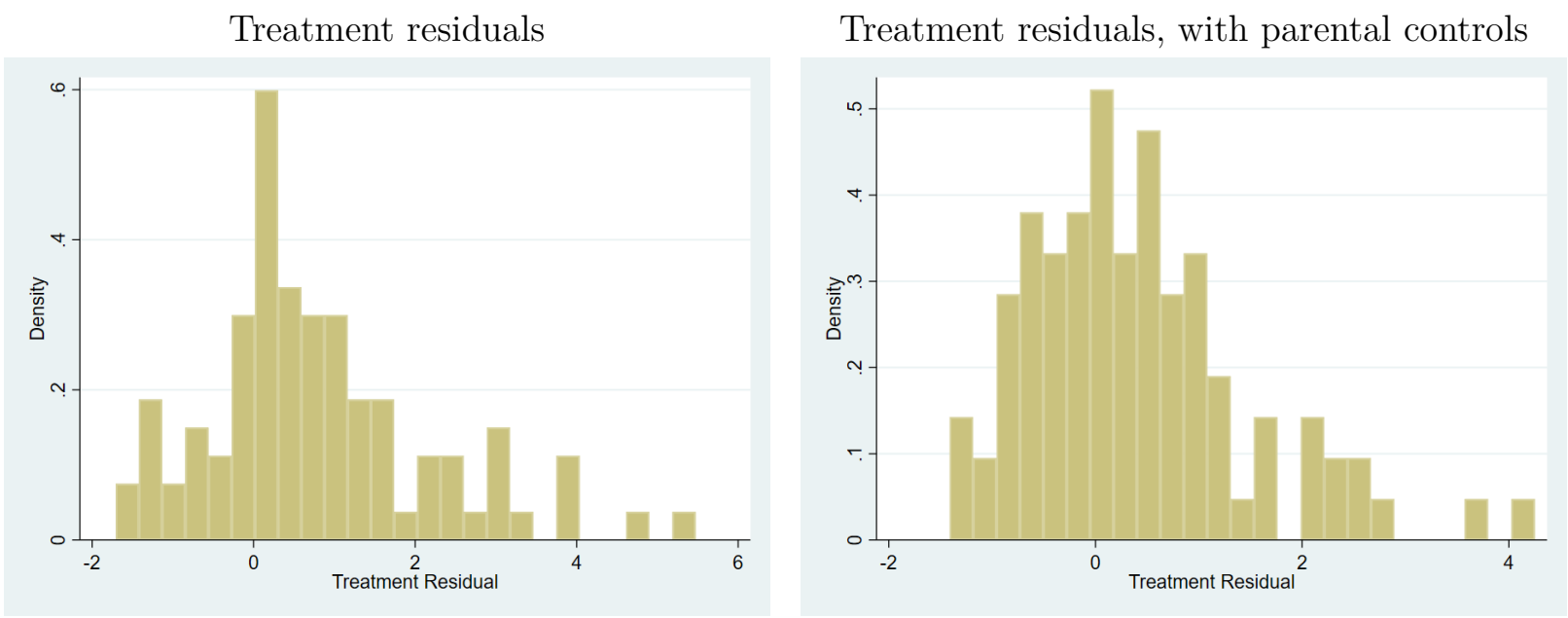
Goodman-Bacon (2021) devises a simple check for implied weights on treatment effects in difference-in-difference estimates, which primarily involves residualizing treatment on the set of controls, consistent with the Frisch-Waugh-Lovell interpretation of regression.

We can perform a similar analysis to inspect our data for the presence of negative weights, and especially of large negative weights, on the outcomes in treated state-cohorts. Figure 3 shows the distribution of the simple accountability measure residualized on controls (including state and cohort fixed effects) among state-cohorts which are assigned more than zero years of accountability by the simple assignment rule. The left half shows values residualizing only on fixed effects, while the right half shows accountability residualized on both fixed effects and parental controls. Negative values correspond with negative weights on treatment effects in a regression analysis, since exposure to treatment is lower than predicted by controls. Figure 3 shows that a limited (though nontrivial) number of treated observations receive negative weight in the baseline specification, while a larger number receive negative weight in the specification including controls.

## C.2 Effects on attainment by year

Table 4 reports estimates of the effect of accountability on attainment by specific grade/degree. For simplicity, we report only results from Specification 1 with and without controls. Each cell is an estimate of the coefficient of interest. Each row represents an outcome variable, while each column represents a choice of inclusion of controls, with the first column not including parental controls and the second column including them.

Figure 3: Residuals




Histogram of residual of simple accountability measure on controls among state-cohorts with at least one year of exposure to accountability. Left: controls are state and cohort fixed effects alone. Right: adding parental controls. Negative values signify observations whose treatment effect receives negative weight in main specifications.

Table 4: Detailed educational attainment results

| | (1) $YrsAcc$ | (2) (with controls) |
|---|---|---|
| At least 10th grade | 0.0010 | 0.0011 |
| | (0.0003) | (0.0003) |
| At least 11th grade | 0.0015 | 0.0015 |
| | (0.0003) | (0.0004) |
| At least HS degree | 0.0020 | 0.0019 |
| | (0.0005) | (0.0005) |
| At least some college | 0.0026 | 0.0001 |
| | (0.0008) | (0.0006) |
| At least associates degree | 0.0021 | 0.0004 |
| | (0.0007) | (0.0005) |
| At least bachelor's degree | 0.0020 | 0.0003 |
| | (0.0008) | (0.0007) |
| Graduate degree | 0.0019 | 0.0008 |
| | (0.0007) | (0.0005) |
| $N$ | 630 | 630 |

Each entry is a coefficient estimate from a separate regression, on $YrsAcc$ ("simple" measure) using Specification 1. Outcome variables are attainment dummies and vary by row. Column 1 controls for state and cohort fixed effects, while column 2 additionally controls for parental characteristics. Standard errors in parentheses clustered by state.

## C.3 Effect of accountability on probability of attending public school and remaining in-state

We interpret our estimates using the detailed accountability measure as estimates of the effects of exposure to accountability itself. However, it is theoretically possible for assignment to accountability to affect long-term outcomes through other channels than directly experiencing accountability. Two alternate mechanisms are that assignment to accountability might have made children more likely to attend private school or to have moved out of state.

To assess these possibilities, we construct a sample of all children attending school in both the 1990 and 2000 5% Census samples. This allows us to observe private school attendance and if a student had moved states both prior to exposure and, for most treated states, after the adoption of a consequential accountability regime. As outcomes, we take the state-Census year average of indicators for (i) if an individual is attending private school and (ii) if they are attending school in their birth state. Our right-hand side variables of interest are (i) whether consequential accountability was introduced between 1990 and 2000 and (ii) the change in the total number of grades tested between 1990 and 2000. For each of the four resulting combinations of outcome and treatment, we regress the outcome measured in 2000 on the treatment, controlling for the lagged outcome (i.e., outcome measured in 1990).

Results are shown in Table 5. Each entry represents a separate specification.

For both definitions of treatment (change in presence of accountability and change in years of accountability), we find zero effect on the percentage of students attending private school. By contrast, we estimate accountability statistically significantly decreased the probability a student remained in their birth state. However, this effect is so small – with roughly 1 in every 500 students moving out of state per year of accountability – that it could not meaningfully affect our main results involving labor market and educational attainment. For example, Chetty and Hendren (2018) estimate that growing up in a one standard deviation better county at producing incomes in adulthood increases earnings in adulthood by 6-10%. So, if each year of exposure to accountability induced 1 in every 500 students to move to a one standard deviation county for their entire childhood, this would increase average incomes by less than 1/50th of a percent. This is an order of magnitude smaller than either our results or the uncertainty in them arising from sampling error.

## C.4 Effect of accountability on probability of missing data

Our results may be biased by sample selection if accountability affects the probability an individual works, and therefore that their occupation is observed. To assess this, we construct as additional outcomes the percentage of observations where we observe an occupation for an individual and a positive income at the state-cohort level, respectively. Table 6 reports the results from our two-way fixed effects specification both with and without controls. Accountability appears not to have affected whether or not an occupation or positive income is observed.

Table 5: Effects of consequential accountability on enrollment

| Change in accountability | (1) Percent in birth state | (2) Percent enrolled in private school |
|---|---|---|
| Consequential accountability introduced | -0.0083 | 0.0027 |
| | (0.0049) | (0.0032) |
| Change in total exposure to accountability | -0.0016 | 0.0005 |
| | (0.0007) | (0.0005) |
| $N$ | 45 | 45 |

Each entry is a coefficient estimate for a separate specification. Rows are independent variables, columns are dependent variables. Dependent variables measured in 2000. All specifications control for the 1990 value of the dependent variable. Observations are a state. Robust standard errors in parentheses.

Table 6: Probability of missing occupation and zero income

| | (1) Fraction with observed occupation | (2) Fraction with no income |
|---|---|---|
| $YrsAcc$ | 0.0000 | 0.0007 |
| | (0.0004) | (0.0006) |
| (with controls) | 0.0009 | -0.0001 |
| | (0.0004) | (0.0005) |
| $N$ | 630 | 630 |

Each entry is a coefficient estimate from a separate regression, on $YrsAcc$ ("simple" measure) using Specification 1. All estimates control for state and cohort fixed effects. Rows labeled "with controls" are estimates additionally including parental controls (log of average income, average educational attainment, and average occupational skills) controlling separately for average values of parents native to the state and parents who migrated. Standard errors clustered by state in parentheses.

## C.5 Robustness to changes in definition of control group

Tables 7 and 8 report results from Specification 1, with and without controls, excluding control observations which might plausibly be considered to have been treated. In particular, the three panels of each table exclude observations where, using the simple assignment rule, the person would not have completed grade 12 at the time that accountability is first implemented (Panel A); the individual lived in a state which ever introduced "report card" accountability, i.e. an accountability policy which we do not rate as consequential, typically involving simply reporting test score information to the public (Panel B); or both of these exclusions (Panel C). Each cell represents a coefficient of interest in that regression.

Table 7: Education results with restricted control groups

| | (1) Years Educ | (2) Complete HS | (3) Complete BA |
|---|---|---|---|
| **Panel A:** Dropping under-regime cohorts | | | |
| *YrsAcc* | 0.0236 | 0.0021 | 0.0024 |
| | (0.0049) | (0.0005) | (0.0009) |
| (with controls) | 0.0160 | 0.0023 | 0.0009 |
| | (0.0046) | (0.0006) | (0.0009) |
| **Panel B:** Dropping report card states | | | |
| Years_Acc | 0.0225 | 0.0021 | 0.0024 |
| | (0.0050) | (0.0005) | (0.0009) |
| (with controls) | 0.0068 | 0.0018 | -0.0000 |
| | (0.0044) | (0.0006) | (0.0009) |
| **Panel C:** Dropping both | | | |
| *YrsAcc* | 0.0248 | 0.0021 | 0.0025 |
| | (0.0051) | (0.0005) | (0.0010) |
| (with controls) | 0.0088 | 0.0020 | 0.0001 |
| | (0.0054) | (0.0007 | (0.0011) |

Each entry is a coefficient estimate from a separate regression, on $YrsAcc$ ("simple" measure) using Specification 1. All estimates control for state and cohort fixed effects. Rows labeled "with controls" are estimates additionally including parental controls (log of average income, average educational attainment, and average occupational skills) controlling separately for average values of parents native to the state and parents who migrated. Panels differ in sample restrictions. Standard errors clustered by state in parentheses.

Table 8: Labor market results with restricted control groups

| | (1) ln(Income) | (2) Writing | (3) Math | (4) Creative | (5) Thinking | (6) Science | (7) Non-tested |
|---|---|---|---|---|---|---|---|
| **Panel A:** Dropping under-regime cohorts | | | | | | | |
| *YrsAcc* | 0.0040 | 0.0048 | 0.0028 | 0.0032 | 0.0038 | 0.0006 | 0.0039 |
| | (0.0023) | (0.0019) | (0.0020) | (0.0018) | (0.0020) | (0.0016) | (0.0016) |
| (with controls) | 0.0060 | 0.0033 | 0.0028 | -0.0003 | 0.0022 | 0.0020 | -0.0007 |
| | (0.0029) | (0.0021) | (0.0018) | (0.0017) | (0.0022) | (0.0018) | (0.0015) |
| **Panel B:** Dropping report card states | | | | | | | |
| *YrsAcc* | 0.0023 | 0.0039 | 0.0018 | 0.0024 | 0.0028 | 0.0003 | 0.0029 |
| | (0.0022) | (0.0022) | (0.0017) | (0.0020) | (0.0020) | (0.0013) | (0.0014) |
| (with controls) | -0.0016 | -0.0008 | -0.0000 | -0.0031 | -0.0022 | -0.0025 | -0.0021 |
| | (0.0019) | (0.0016) | (0.0015) | (0.0014) | (0.0013) | (0.0016) | (0.0016) |
| **Panel C:** Dropping both | | | | | | | |
| *YrsAcc* | 0.0037 | 0.0054 | 0.0042 | 0.0032 | 0.0043 | 0.0012 | 0.0046 |
| | (0.0025) | (0.0022) | (0.0022) | (0.0020) | (0.0022) | (0.0018) | (0.0017) |
| (with controls) | -0.0008 | -0.0002 | 0.0020 | -0.0027 | -0.0016 | -0.0009 | -0.0004 |
| | (0.0023) | (0.0023) | (0.0021) | (0.0020) | (0.0021) | (0.0020) | (0.0017) |

Each entry is a coefficient estimate from a separate regression, on $YrsAcc$ ("simple" measure) using Specification 1. All estimates control for state and cohort fixed effects. Rows labeled "with controls" are estimates additionally including parental controls (log of average income, average educational attainment, and average occupational skills) controlling separately for average values of parents native to the state and parents who migrated. Panels differ in sample restrictions. Standard errors clustered by state in parentheses.