AERA Open
July-September 2017, Vol. 3, No. 3, pp. 1–21
DOI: 10.1177/2332858417726324

The Effects of No Child Left Behind on Children's Socioemotional Outcomes

Camille R. Whitney
Independent Education Researcher
Christopher A. Candelaria
Vanderbilt University

Many people have worried about possible adverse effects of high-stakes testing on socioemotional outcomes. This article uses a difference-in-differences approach to investigate the effects of the introduction of high-stakes testing via the No Child Left Behind Act on socioemotional outcomes. Data are from the Early Childhood Longitudinal Survey–Kindergarten Cohort of 1998–1999, a nationally representative longitudinal survey. The 10 outcomes that we examine are from the children's Self-Descriptive Questionnaire, including externalizing problems related to paying attention and behavior in school, internalizing problems related to feeling sad and lonely and academic anxiety, as well as interest and competence in math, reading, and school in general. We find that the introduction of high-stakes test accountability did not have consistent significant effects on these socioemotional outcomes. These findings can help states address concerns and motivate further research on potential unintended consequences of revised accountability systems under the Every Student Succeeds Act.

Keywords: *NCLB, socioemotional outcomes, accountability*

The No Child Left Behind (NCLB) Act of 2001 was the first national law to require consequences for U.S. schools based on students' standardized test scores. Although the NCLB era officially came to a close in December 2015, the Every Student Succeeds Act (ESSA), NCLB's replacement, continues to include consequences for schools according to standardized test scores. Unlike NCLB, ESSA allows greater flexibility in implementation and requires states to choose another measure of school quality beyond four required academic indicators. Examining the benefits and costs of consequential accountability systems under NCLB can help inform current state efforts to revise their accountability systems under ESSA. This article examines effects of NCLB on socioemotional outcomes, a topic that has generated much speculation but little research to date.

Critics charge that high-stakes accountability systems such as NCLB exact a heavy cost on students' socioemotional well-being. For example, a greater focus on testing could cause higher levels of anxiety or test anxiety among children (M. G. Jones et al., 1999; Madaus, 1991; Paris, Lawton, Turner, & Roth, 1991; Segool, Carlson, Goforth, Von Der Embse, & Barterian, 2013; Wolf & Smith, 1995). However, few rigorous large-scale quantitative studies have examined whether high-stakes standardized testing affects socioemotional outcomes.

In this article, we aim to address this gap in the literature by measuring the impact of NCLB on a variety of socioemotional outcomes, using the Early Childhood Longitudinal Survey–Kindergarten Cohort of 1998–1999 (ECLS-K). We refer to such testing as "consequential accountability," meaning that failing to meet certain statewide standards has serious consequences for schools, as defined by Hanushek and Raymond (2005). Using a difference-in-differences quasi-experimental methodology, we address the following research questions:

Research Question 1: What effects did the introduction of consequential accountability via NCLB have on children's socioemotional outcomes?
Research Question 2: Are there different effects for different student subgroups as defined by socioeconomic status and gender?

The outcomes include student-reported externalizing behaviors (misbehaving at school and failing to pay attention), internalizing behaviors (academic anxiety and feeling sad or lonely), and interest and self-competence in math, reading, and all school subjects in general. For our purposes, we refer to all of these as "socioemotional" outcomes. We choose this term because each outcome involves how students feel (internalizing, interest, self-competence) or how they act in a social setting (externalizing), but some of the outcomes do not fit common definitions of "socioemotional" outcomes. These types of outcomes have also been called

"noncognitive factors," although that phrase is problematic given that the outcomes involve cognition (Farrington et al., 2012). Like Farrington et al. (2012), we do not use the term "competencies" or "skills" but instead use "outcomes" since these measures encompass competencies and well-being.

Our results suggest that NCLB consequential accountability did not have a consistent discernible effect on the 10 socioemotional outcomes that we examined. While we find a few significant point estimates suggesting a modest increase in academic anxiety and marginally significant increases in math interest and sense of competence, these results may have occurred by chance given the number of significance tests that we perform. Overall, the results are useful in that they can inform policy makers and school leaders about specific socioemotional outcomes that do not appear to be affected, on average, by high-stakes testing regimes coupled with consequential accountability. From a policy perspective, this knowledge can steer further discussion and research in a more nuanced direction, toward investigating and mitigating potential negative responses associated with consequential accountability among subgroups of schools or students while promoting positive aspects.

We also perform auxiliary exploratory subgroup analyses. We find suggestive evidence that high-stakes standardized testing produced an increase in academic anxiety for students in the top half of the income distribution. We also find evidence that students in the bottom half of the distribution of socioeconomic status experienced increases in their competence and interest in math. Given these subgroup results are exploratory, the causal warrant and generalizability of these estimates are limited.

Background

The NCLB Act, signed into law in January 2002, ensured that consequential accountability would be implemented on a national scale. States were required to have annual tests at every school in Grades 3 to 8 in math and reading, as well as at least once during high school.[1] Schools were required to meet an annual performance target for the state, adequate yearly progress (AYP), with regard to test scores and sufficient participation rates. These AYP targets were set relative to performance at baseline, in spring of the 2001–2002 school year. Furthermore, schools needed to make AYP not only for their student bodies overall but also for each student demographic subgroup (e.g., race, disability status) that was large enough to be considered a significant subgroup at that school, according to the state's threshold. Schools that failed to make AYP faced sanctions that grew more severe each year that they failed.[2] Generally, the first year of failure put the school on probation, and in subsequent years, the school faced escalating sanctions, such as mandatory participation in school improvement programs. Ultimately, a failing school required closure, restructuring, or "any other major changes in school governance." NCLB included other important provisions as well, but only the consequential accountability piece is studied in this article, as described in the Methodology section. This article focuses on the effects of the high-stakes testing provision of NCLB on socioemotional outcomes of NCLB in the early years of the policy only.

Conceptual Framework and Related Literature

Standardized tests and associated accountability measures may affect school staff members' behaviors and attitudes, which in turn may affect students' behaviors, anxiety and internalizing, and interest and self-competence in school subjects. In what follows, we draw on theories and empirical results from prior literature to help describe possible mechanisms by which such policies could affect students' socioemotional well-being. This framework demonstrates that the expected directions of effects of NCLB on socioemotional outcomes are unclear—they could be positive, negative, or null—and they might differ by student subgroups.

Possible Negative Effects

Theory, as well as some empirical evidence, suggests that changes in school behaviors as a result of consequential accountability could be detrimental to students' socioemotional well-being. Theories from psychology imply that high-stakes tests could lead to greater anxiety. Self-affirmation theory says that people feel a need to maintain a positive self-image to feel secure and satisfied (Steele, 1988). Thus, a high-stakes test might cause a student to be anxious about scoring well enough to protect one's self-image. Psychology research also finds that when students have a performance goal, such as doing well on a test, they tend to feel more anxious than when focusing more on process goals (Ryan & Ryan, 2005). In turn, increased anxiety about tests and school performance might compromise children's learning and well-being. Hundreds of studies have linked test anxiety (Hembree, 1988) and other types of anxiety (Seipp, 1991) with decreased academic performance, and internalizing puts children at risk of learning problems, lower achievement, and lower social skills (e.g., problem solving; Kovacs & Devlin, 1998).

Consequential accountability might have negative effects on children's interest in school subjects on average, which might lead to worse behavior. Students' overall school interest might decrease if schools spend more instructional time on math and reading and less on other subjects or activities that many children find interesting. Such instructional time shifts have been observed due to NCLB and other performance-based accountability systems (Hannaway & Hamilton, 2008). Dee, Jacob, and Schwartz (2013) find that NCLB caused a narrowing of the curriculum in terms of a greater proportion of instructional time spent on tested subjects,

although the effect was relatively small in size. Another study found that 44% of school districts said that they had cut time from other subjects and activities so that students would have more instructional time in math and reading after NCLB (McMurrer, 2007).

Tested subjects might become less interesting for students, perhaps due to teachers "teaching to the test" in a rote manner, which in turn could lead to more externalizing behaviors among children acting out because of boredom. One study finds that third-grade students are less engaged during "large group activities, individualized work settings, and basic skills instruction" as compared with small group activities and instruction emphasizing analysis and inference (Downer, Rimm-Kaufman, & Pianta, 2007), suggesting that instructional practices commonly associated with test preparation are less engaging for students, although there is no conclusive evidence that teachers are engaging more in these behaviors as a result of NCLB.

Possible Positive Effects

While there is little direct evidence that instruction improved as a result of consequential accountability under NCLB, we might infer that school instruction and/or the learning environment has improved given findings of positive effects on test scores. Based on quasi-experimental methods, several studies have found that strong accountability yields some positive gains in achievement, although more consistently in math than in reading (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Wong, Cook, & Steiner, 2009, 2015). In addition, Hannaway and Hamilton (2008) summarize a number of studies that found that some teachers work harder in response to consequential accountability and have a greater focus on and capacity for good instruction. Any improvements in instruction due to NCLB would likely have also translated to improved student interest and behavior.

One study, based on quasi-experimental methods, finds only positive or null associations with socioemotional outcomes when schools face particularly high NCLB accountability pressure (Reback, Rockoff, & Schwartz, 2014). The authors exploit the fact that rules related to the requirements for making AYP under NCLB vary across states. For example, each state can determine how many students a school must have in a subgroup to make it count as "significant" and thus count toward the school's AYP determination. Because of these differences in rules, a school that is at risk of failing AYP in one state could be expected to be quite far above the AYP cutoff in another state. The authors assume that the schools that are near the margin of failing feel more test pressure than those well above (or below) the margin, and so they can compare similar schools across these "higher-pressure" and "lower-pressure conditions" to identify the possible effect of facing more test pressure on student outcomes. Using selected items from the fifth-grade ECLS-K, the authors found a statistically significant decrease in the single item that measured test anxiety and no significant effect on indexes that they created of "enjoyment" of reading and math, with four items each, although point estimates are in a positive direction (Reback et al., 2014).[3]

One other quasi-experimental study considers the effects of NCLB on socioemotional outcomes (Dee et al., 2013). Using multiple waves of the Schools and Staffing Survey, the authors create an index of teachers' perceptions of students' behavioral engagement. Items in the index measure apathy, tardiness, class cutting, absenteeism, coming to class unprepared, and causing problems for the school. The study finds that teachers' perceptions of student behavioral engagement increased as a result of NCLB through 2007–2008, although the results are reduced and no longer significant with a traditional difference-in-difference approach rather than comparative interrupted time-series design.

Possible Differential Effects

In this section, we explain that some differences in effects might come from systematic differences in the type of school or labor market experienced by students in these subgroups as a result of NCLB. Students might have different experiences due to segregation across schools or due to differences in teacher behaviors across types of students within the classroom. In addition, different types of students might react differently to the same types of NCLB-related changes due to factors such as stereotype threat.

Different student subgroups might have different experiences of NCLB due to between-school sorting. For example, students of lower socioeconomic status are more likely to be clustered in schools in danger of failing to make AYP. These students might therefore experience stronger responses from their schools to NCLB and thus larger treatment effects, positive or negative, than other students. Empirical evidence finds that schools do in fact respond differently to NCLB according to their student population. For example, one study finds that urban and high-needs schools tended to lower the cognitive demand of their English language arts classes in the years following NCLB, while suburban or lower-needs schools did not lower cognitive demand (Polikoff & Struthers, 2013).

Studies have found strong effects of standardized tests on the labor market that appear differentially negative for disadvantaged students who are clustered in schools that serve predominantly disadvantaged students. Teachers are more likely to leave schools that have recently been given a negative "accountability shock" and are less likely leave schools that have experienced a positive shock, corresponding with an increase in teacher quality in more versus less advantaged schools (Feng, Figlio, & Sass, 2010). Principals of higher quality also move to less disadvantaged schools that are less

likely to fail AYP, causing average quality of principals to fall in more disadvantaged schools (Li, 2012). If higher-quality principals and teachers in terms of achievement are also better than others at fostering their students' socioemotional well-being, then these changes in the distribution of teachers and principals could hurt more disadvantaged students relative to more advantaged students.

Within schools, students from different subgroups may have different experiences as well. For example, teachers might treat students from different subgroups differently. Teachers might have biases regarding certain subgroups of students, or they might act strategically in response to NCLB by focusing their efforts on students likely to score near the cutoff for proficiency, known as the "bubble kids." Research has found that scores rise more among students who are likely to determine their school's accountability (Reback, 2008), and test score gains from NCLB have been concentrated among students in the middle of the test score distribution (Neal & Schanzenbach, 2010). However, Reback et al. (2014) found that students in a subgroup that placed their school in danger of failing AYP fared equally well in terms of test anxiety and "enjoyment" in math and reading as students outside that subgroup.

In addition to different experiences by subgroup, different responses to test pressure across subgroups could cause differences in effects. For example, if tests do increase stress among children, research finds that boys and girls react to stress differently, with girls more likely to exhibit internalizing symptoms (Leadbeater, Blatt, & Quinlan, 1995). We might expect different effects on socioemotional outcomes across student subgroups given that effects on achievement test results vary by subgroup, with less advantaged groups generally showing more positive results. For example, Dee and Jacob (2011) find that fourth grade math test score gains were especially large among Black and Hispanic students, those eligible for free lunch, and students in the lowest decile of achievement. They do not find differences by gender, though. These findings suggest the importance of studying subgroup-specific effects, although it is not clear whether groups that experience differentially positive achievement effects should also be expected to experience particularly good effects on socioemotional outcomes, especially since the mechanisms behind the achievement gains are far from clear.

Summary and Contribution

While the literature clearly indicates that schools change their behaviors in response to consequential accountability, it also indicates that children might show positive or negative changes in their socioemotional outcomes as a result. We could see negative effects if students experience greater test-based pressure or if students are less engaged in school due to rote teaching to the test or narrowing of the curriculum. We could see positive effects from mechanisms such as improvements in instruction or greater emotional support for students. The positive story is supported by the findings of lower test anxiety from Reback et al. (2014) and higher teacher-rated engagement from Dee et al. (2013).

This article contributes to the literature by addressing a gap in our knowledge of the effect of high-stakes testing on students' socioemotional outcomes, including socioemotional outcomes that have not been examined in prior causal studies of the effect of NCLB. For example, while Reback et al. (2014) examine "test anxiety" as an outcome of interest, we examine "academic anxiety," a more comprehensive school-based anxiety measure that we constructed from an exploratory factor analysis. In addition, this article contributes new findings based on a quasi-experimental methodology that leverages the longitudinal nature of the ECLS-K data to estimate the average treatment effect of the introduction of high-stakes testing; specifically, we assess whether students' socioemotional outcomes were affected by the introduction of NCLB high-stakes testing between the third and fifth grade. In their related study, Reback et al. use a differences-in-differences approach that relies on comparisons of schools that are below, on, or above the AYP margin in the spring of 2004, when students are in the fifth grade, to estimate effects on socioemotional outcomes. Consequently, our results are complementary to those of Reback et al. Finally, this study examines heterogeneous effects of NCLB high-stakes testing by examining possible differential effects on student subgroups.

Data

The data for this study come from the ECLS-K. Data were collected in the fall and spring of the kindergarten year, spring of the first grade year (with a fall subsample), and spring of the third-, fifth-, and eighth-grade years. The data set is nationally representative of kindergartners in 1998–1999. This study uses data from the spring third- and fifth-grade years, although we use other years to construct some of our baseline covariates. Note that students were assessed around the same time across schools (95% were assessed in March, April, or May in both the third- and fifth-grade years—the same time of year as when students take high-stakes standardized tests, meaning that high-stakes testing should have been salient for students at that time).

This data set offers important advantages for addressing this article's research questions. It is one of the few data sets to include socioemotional outcomes at a national level. With respect to our difference-in-differences strategy, the causal warrant is stronger, as we can follow the same set of children from one year to the next. In addition, the survey had fortuitous timing with regard to NCLB. The third-grade round occurred before NCLB consequences were put into place: the law was signed in January 2002, in the middle of the third-grade year. During the spring of that year, students

took tests that established baseline scores for judging schools' performance in subsequent years. Schools could fail to make AYP based on their test scores in the next year, 2002–2003 (i.e., the students' fourth-grade year). Thus, by the time that students were surveyed during the fifth-grade year in the spring of 2004, NCLB consequences were widely in effect. This provides an ideal setup for the type of difference-in-differences approach that we describe here.

Outcome Variables

Socioemotional outcome variables come from the children's Self-Descriptive Questionnaire (SDQ), which is a reliable and valid instrument according to field testing (Atkins-Burnett & Meisels, 2001; Pollack, Atkins-Burnett, Najarian, & Rock, 2005). Because children were given exactly the same survey in the spring of their third- and fifth-grade years, this instrument provides an ideal opportunity to consistently examine NCLB outcomes, before and after its implementation. In what follows, we briefly describe the types of items that compose our subscales; however, we provide the exact wording of the items with additional details of the SDQ in Appendix A.

Among the six available SDQ-based scales included in the ECLS-K, we selected five for this analysis; the sixth—Perceived Interest/Competence in Peer Relations—seemed less likely to be affected by high-stakes testing. For each item, students could give one of four responses: *not at all true*, *a little bit true*, *mostly true*, or *very true*, with corresponding point values of 1 through 4.

We conducted exploratory factor analysis, which suggested two factors per scale from the SDQ. These factors are the outcome subscales that we use. For each outcome subscale, a child's score is a simple mean of the items. Additional details about the factor analysis appear in Appendix A. The appendix includes technical details about the factor analysis, the list of items for each outcome subscale, and Cronbach's reliability coefficient for each constructed subscale, which ranges from 0.62 to 0.92.

Externalizing problem behaviors yielded one factor measuring difficulty paying attention (e.g., "I get distracted easily"), which is composed of three items. The other factor measured student misbehaviors (e.g., "I get in trouble for talking and disturbing others"), and it comprises two items.

Factor analysis of the Internalizing Problem Behaviors Scale yielded one subscale with four items relating to academic anxiety (e.g., "I worry about doing well in school"), including an item that asks about test anxiety: "I worry about taking tests." The other subscale includes two items about feeling sad or lonely.

The other three scales relate to interest and self-competence in reading, math, and all school subjects, respectively. Many of the interest items are quite straightforward in asking whether the student likes the subject (e.g., "I like reading"), and self-competence items ask, for example, whether the student does well in the subject and gets good grades. We used factor analysis to create outcome scales composed of the interest and competence items for each of the three scales. The reading, math, and "all school subjects" interest subscales comprise five, four, and three items, respectively, while the reading, math, and "all school subjects" competence subscales comprise three, four, and three items, respectively.

Methodology

This article implements a difference-in-differences strategy to identify the effect of NCLB consequential accountability on socioemotional outcomes. Building off the work of Hanushek and Raymond (2005) and Dee and Jacob (2011), we exploit the fact that some states already had accountability with strong consequences before NCLB began. We would not expect NCLB to cause a change in socioemotional outcomes in these states, since NCLB would not be initiating greater test pressure on schools in these states or, in other words, the schools' production functions would already have shifted toward more emphasis on test score outputs prior to NCLB. Other states, however, did not have such consequential accountability systems in place before NCLB. While most of these states had statewide tests in at least some grades, consequences for most schools in these states were considered weak or null. This analytic strategy assumes that schools in these states experienced a much greater increase in test pressure after NCLB went into place. These are the "treatment" states, while the states with prior consequential accountability are the "comparison" states. The analysis relies on an assumption that changes in comparison states serve as a counterfactual for the changes that we would have observed in treatment states in the absence of NCLB, since these changes should be caused by non-NCLB factors, such as other school reforms or shifts in the economy. The methodology, assumptions, and the specific model used for analysis are discussed in more detail.

To identify the effect of the introduction of high-stakes testing on socioemotional outcomes, we posit the following estimation equation:

$$Y_{ist} = \mu_s + \beta_1 Post_t + \beta_2 (T_s \times Post_t) + \mathbf{X}_{it}'\gamma + \varepsilon_{ist},$$

where Y_{ist} is the socioemotional outcome of interest for student i in state s in year t, standardized within year t; $Post_t$ is a binary indicator variable that takes value 1 in the year 2003–2004 and is equal to 0 in the year 2001–2002; T_s is a binary indicator variable that takes value 1 if a student resides in a state that did not have prior consequential accountability and 0 otherwise; μ_s is a state-specific fixed effect; $\mathbf{X}_{it}$ is a vector of time-varying and time-invariant covariates; and ε_{ist} is a mean-zero random error term.

Standard errors are clustered at the state level—the level at which treatment occurs—to address concerns with serial correlation in the error term over time (Bertrand, Duflo, & Mullainathan, 2004). When estimates are weighted, we use the third- and fifth-grade longitudinal weight, which weights the sample to be nationally representative of all children who entered kindergarten in 1998–1999. The parameter of interest is β_2, which gives the causal effect of the introduction of consequential accountability via NCLB. Note that other provisions of NCLB, such as those requiring highly qualified teachers, are not measured by β_2, since these provisions affected the treatment and comparison states and should therefore be subsumed in the common trend over time, β_1.

The causal interpretation attached to β_2 relies on certain assumptions. One must assume that consequential accountability policies in the comparison states (i.e., those enacted pre-NCLB) were similar in terms of their impacts on socioemotional outcomes to those that were implemented during the NCLB era. If not, then students in the comparison states are not truly "controls" in the sense that they might also experience a change in their socioemotional outcomes as a result of NCLB. In fact, NCLB may have increased pressure in comparison states on average, to the extent that its consequences were stronger than what was in place before in some of these states, including, in some cases, more stringent subgroup-specific accountability than before (Dee & Jacob, 2011). The literature described finds that schools have often responded by making changes in their behavior when faced with consequential accountability prior to NCLB (e.g., Boyd, Lankford, Loeb, & Wyckoff, 2008; Jacob & Levitt, 2003). Thus, schools in these states have likely already experienced some sort of increased pressure and responded to it; therefore, it seems logical that NCLB would have a stronger effect in states where these responses have not yet been prompted. A related concern is that the treatment states are likely more reluctant adopters of consequential accountability than the comparison states and thus might have experienced less of an impact than would the average state (Dee & Jacob, 2010). These issues would attenuate results from this article's analysis.

One additional concern is that students in the treatment states were already feeling some effects of consequential accountability when they were surveyed in third grade. Since NCLB was signed in January 2002 and third-grade students were surveyed a few months afterward, schools might have already been making some shifts toward preparing for consequential accountability testing in the coming years. Still, we would expect the shifts to be much greater once NCLB actually began. If treatment state schools were making some changes already in third grade, again this would attenuate any results in the analysis, leading to an underestimation of the effects of consequential accountability on socioemotional learning (SEL) outcomes.

This analytic strategy assumes that nothing else happened between the time that socioemotional outcomes were measured in third and fifth grade that affected these measures differently in treatment versus comparison states except for NCLB. For example, it could be that the composition of public school students changed in treatment states because certain parents placed their children in private schools in treatment states. Dee and Jacob (2011) perform a number of auxiliary regressions to check for such changes. They find only a very small shift from public to private schools due to NCLB (about 1% of students), and the authors find little evidence that NCLB caused substantial shifts in student composition. They also find that other important factors, such as poverty rates, median household income, and employment, did not shift as a result of NCLB. Thus, the Dee and Jacob article lends credence to the assumption that treatment effects were caused by NCLB and not by something else that affected treatment states differently from comparison states.

Analytic Sample Description

Table 1 reports weighted means of student and school characteristics in Grade 3. In total, approximately 7,950 students are included, meaning that these students had a third- and fifth-grade school ID number that could be linked to a treatment or comparison state, attended public school in the same state in third and fifth grade, and had outcome measures and covariates. In supplementary analyses (not shown), the original sample and the analysis samples appear to be nearly identical in terms of average values on observable characteristics, assuaging concerns that the sample for the analysis might not be widely generalizable.

Because a majority of states implemented consequential accountability prior to NCLB, the sample is unbalanced between treatment and comparison states and students. For the student-reported items, students number about 2,310 in the 15 treatment states and 5,640 in 25 comparison states. Table 2 lists the states in each treatment condition. When we test for differences between the two groups and adjust for serial correlation at the state level, we do not find any statistically significant differences between treatment and control means.

The variables in Table 1 also serve as controls in our regression specifications. When analyses use control variables, the time-invariant covariates are the student-level indicators for race/ethnicity, gender, socioeconomic status, and first-grade *t* scores in math and reading. The socioeconomic status variable (students composite) combines information on parents' income, education level, and occupational prestige, and we average this variable between kindergarten and third grade. Time-varying covariates include school enrollment, proportion Hispanic, proportion Black, and the proportion of students who are eligible for free and reduced-price lunch. In addition, we include an indicator for whether

TABLE 1
Summary Statistics: Weighted Means

	Full sample	Treatment	Control
Student characteristics			
Demographics			
Proportion Black	0.18	0.11	0.21
Proportion Hispanic	0.18	0.12	0.21
Proportion Asian	0.03	0.02	0.03
Proportion other race (non-White)	0.04	0.08	0.03
Proportion male	0.52	0.51	0.52
SES composite (K–3)	−0.14	−0.06	−0.17
First-grade *t* scores			
Reading	49.69	50.62	49.35
Math	49.82	50.91	49.42
School characteristics			
Enrollment	5.46	4.64	5.75
Proportion Hispanic	0.16	0.09	0.18
Proportion Black	0.19	0.14	0.21
Proportion FRPL eligible	0.46	0.39	0.49
No. of states	40	15	25
No. of children	7,950	2,310	5,640

Note. Number of children is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. All means are computed according to baseline data. Mean estimates are weighted with a third- through fifth-grade longitudinal weight to achieve national representation. The socioeconomic status composite variable is averaged between kindergarten and third grade. School characteristics data are a combination of school-level data from the Common Core of Data School Universe files as well as the Early Childhood Longitudinal Survey–Kindergarten files. SES = socioeconomic status; FRPL = free and reduced-price lunch.

TABLE 2
Treatment and Control States

Treatment states	Control states
(No pre-NCLB consequential accountability)	(Pre-NCLB consequential accountability)
Arizona	Alabama
Colorado	Alaska
Hawaii	California
Iowa	Connecticut
Maine	Delaware
Minnesota	Florida
Mississippi	Georgia
Missouri	Illinois
New Jersey	Indiana
Ohio	Kansas
Pennsylvania	Kentucky
South Dakota	Louisiana
Utah	Maryland
Washington	Massachusetts
Wyoming	Michigan
	New Mexico
	New York
	North Carolina
	Oklahoma
	Oregon
	Rhode Island
	Tennessee
	Texas
	Virginia
	Wisconsin

Note. As defined by Dee and Jacob (2011), control states adopted consequential accountability prior to No Child Left Behind (NCLB); treatment states had no prior consequential accountability. Our lists differ from Dee and Jacob's in that Arkansas, Nevada, South Carolina, and West Virginia are excluded from the control group and Idaho, Montana, Nebraska, New Hampshire, North Dakota, South Carolina, and Vermont are excluded from the treatment group, as these states are not represented in the Early Childhood Longitudinal Survey–Kindergarten data set.

the school is a charter school and an indicator for whether the child switched schools between third and fifth grade.

Results

We organize results into three sections. The first section reports estimates of the effect that NCLB had on externalizing and internalizing outcomes. The second section reports results that assess the effect of NCLB on interest and competence in school subjects. The third section explores whether there is heterogeneity in the estimated NCLB treatment effects that differ by either socioeconomic status or gender.

Regression tables in the first two sections report estimated NCLB treatment effects for each outcome of interest and examine the robustness of the results. To assess robustness, we perturb our main specification by altering the use of covariates (i.e., including or excluding them), using either state fixed effects or child fixed effects and including or excluding longitudinal population weights. Covariates are primarily included to increase precision in our estimates and to help assuage concerns about omitted variable bias; overall, we find that point estimates are very similar whether controls are in or out of the model. The use of state-level fixed effects is motivated by our difference-in-differences quasi-experimental design; however, by leveraging the variation within each student and controlling for secular time effects, we can identify the NCLB treatment effect while controlling for a host of time-invariant omitted variables that might bias our estimates. Indeed, we find that estimates are similar when comparing state- versus child-level fixed effect models, but using child fixed effects substantially reduces the precision of the estimates. Finally, we estimate models that use and exclude the longitudinal population sampling weight between third and fifth grade. Because our regression analyses do not use the full sample of data for which the weights are originally designed (e.g., we limit our analytic sample to

TABLE 3
Externalizing and Internalizing Results

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
	Panel A: Unweighted regressions			
State fixed effects with covariates				
$T_s \times NCLB_t$	0.011	−0.002	−0.029	0.079*
	(0.023)	(0.037)	(0.032)	(0.037)
State fixed effects, no covariates				
$T_s \times NCLB_t$	0.009	−0.004	−0.029	0.076*
	(0.024)	(0.039)	(0.032)	(0.037)
Child fixed effects with covariates				
$T_s \times NCLB_t$	0.012	0.005	−0.023	0.078
	(0.033)	(0.052)	(0.042)	(0.052)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	0.008	0.005	−0.024	0.079
	(0.033)	(0.053)	(0.043)	(0.053)
	Panel B: Weighted regressions			
State fixed effects with covariates				
$T_s \times NCLB_t$	−0.004	−0.049	−0.051	0.138+
	(0.049)	(0.062)	(0.055)	(0.076)
State fixed effects, no covariates				
$T_s \times NCLB_t$	−0.010	−0.058	−0.055	0.132+
	(0.050)	(0.064)	(0.057)	(0.076)
Child fixed effects with covariates				
$T_s \times NCLB_t$	−0.001	−0.026	−0.043	0.145
	(0.068)	(0.081)	(0.077)	(0.103)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	−0.007	−0.030	−0.044	0.141
	(0.068)	(0.081)	(0.078)	(0.106)
Observations	15,890	15,890	15,890	15,890

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{+}p < .10$. $^{*}p < .05$.

public schools), weighting the data is not necessarily preferred; consequently, if unweighted and weighted results diverge, we use the results to bound the estimate.

The subgroup results focus on specifications that do not include covariates or weights. Covariates are excluded as the goal is to examine heterogeneity across socioeconomic status and gender groups; we do not want to "control" for these characteristics. The decision not to weight stems from the fact that weights cannot provide national representation when examining a given subgroup.

Externalizing and Internalizing Outcomes

Table 3 reports results for externalizing and internalizing outcomes. For externalizing outcomes, the Attention subscale refers to difficulty in paying attention, and the subscale Behavior refers to misbehavior in class. For internalizing outcomes, the Sad/Lonely subscale reflects a student's feelings about sadness or loneliness at school, and the Academic Anxiety subscale reflects worry about performance in school and shame in making mistakes. Increases in any of these subscales suggest that students are faring worse on the dimension under consideration.

Overall, the results suggest that NCLB had a moderate effect on academic anxiety. According to our unweighted regressions, academic anxiety increased by 0.08 standard deviations, which is significant at the 5% level. The weighted estimates suggest that it increased by 0.14 standard deviations, which is significant at the 10% level. Given sample restrictions imposed by missing covariates and limiting the sample to public schools, using the population weights does not necessarily provide national representation; therefore,

we use our unweighted and weighted results to bound the estimates, which implies that academic anxiety increases between 0.08 and 0.14 standard deviations.

These academic anxiety findings should be treated with caution, however. When we perform simulations to gauge statistical power in the presence of multiple-hypothesis testing, we find weak evidence in favor of an effect of NCLB on academic anxiety. In our simulation exercise, we run 5,000 simulations where we randomly assign treatment to states. For each simulation, we estimate our eight regression models ([state fixed effects, child fixed effects] × [covariates, no covariates] × [weighted, unweighted]) for each of the 10 socioemotional outcomes that we examine in this study. After the simulations are completed, we calculate the percentage of times that the smallest *p* value on any of the 10 treatment indictators is less than the *p* value from the corresponding academic anxiety regression in Table 3; we do this for each regression model. Overall, the percentages that we calculate across the eight regression models are rather high, ranging from 43% to 63%, which suggests that the statistically significant academic anxiety results likely occurred by chance. Additional details of the simulation exercise appear in Appendix B.

Each of the externalizing subscales and the remaining Internalizing Sad/Lonely subscale suggest that the NCLB treatment effect is not as precisely estimated: Across all specifications, standard errors are larger than all point estimates, although we can still generally rule out any effects >0.1 standard deviations. The magnitude of these estimates in absolute value is also small.

Interest and Competence Outcomes

Table 4 displays interest and competence subscales for math and reading as well as a category defined as "all school subjects." The comprehensive "all school subjects" category includes math and reading as well as any other subjects students take in the third and fifth grade; we refer to this category as "school." Increases in any of these subscales suggest that students have more interest and competence in the subject or subjects being considered.

Despite imprecision, there is a clear pattern of increased interest in math, reading, and all school subjects after the introduction of NCLB. NCLB appears to have increased math competence by a small amount, but results are only marginally significant at the 10% level. By viewing the unweighted estimates with state fixed effects as a lower bound, NCLB increased self-evaluation of math competence by 0.06 standard deviations. The corresponding weighted regressions provides an upper bound of 0.07 standard deviations. With respect to reading competence, however, the unweighted regressions suggest a decrease, while the weighted regressions suggest an increase, but these effect sizes are <0.03 standard deviations in absolute value. Estimates are in a positive direction for self-evaluation of competence in all school subjects.

Exploratory Subgroup Analyses

When we estimate our primary specification, we obtain an average treatment effect of NCLB on SEL outcomes; however, these estimates potentially mask differential treatment responses by subgroups. As hypothesized in our conceptual framework, the direction of these subgroup effects could be positive or negative, so it is an empirical question that we explore with our difference-in-differences estimation strategy. We display externalizing and internalizing results in Table 5 and interest and competence in school subjects in Table 6. As previously mentioned, all reported results are unweighted and do not include covariates.

With respect to externalizing and internalizing outcomes, the results in Table 5 reveal that most of the point estimates are small and insignificant. The only point estimates that are greater than 0.05 standard deviations are those for academic anxiety. Academic anxiety increased by about 0.07 standard deviations (statistically significant at the 5% level) for those in the top half of the distribution of socioeconomic status and by 0.08 standard deviations (not statistically significant) in the bottom half. With respect to gender subgroups, NCLB increased academic anxiety for males by approximately 0.08 standard deviations (marginally significant at the 10% level). The corresponding point estimate for females is 0.06 standard deviations (not statistically significant). The magnitude of these estimates is close to the unweighted estimate of 0.08 standard deviations from the main analytic sample in Table 3.

By examining subject interest and competence outcomes, the results in Table 6 suggest that there are differential effects by socioeconomic status subgroups in terms of math interest and competence. With respect to math, the results indicate that NCLB caused a 0.09–standard deviation increase in interest and a 0.09–standard deviation increase in competence among those in the bottom half of the socioeconomic status distribution. The corresponding estimates for the top half of the distribution are 0.001 standard deviations for interest and 0.03 standard deviations for competence, neither of which is significant. According to the gender subgroups, there is some evidence that males had a 0.07–standard deviation increase in math interest as a result of NCLB and that females had a 0.07–standard deviation increase in competence, both of which are significant at the 10% level. All estimates for reading and the "all school subjects" group are not statistically significant.

Discussion and Conclusion

Overall, we do not observe consistent effects of NCLB on socioemotional outcomes. Our main results suggest that NCLB caused a moderate increase in academic anxiety (between 0.08 and 0.14 standard deviations) in the early years after it was implemented and that it may have improved math interest and competence particularly among

TABLE 4
Subject Interest and Competence Results

	Math		Reading		School	
	(1) Interest	(2) Competence	(3) Interest	(4) Competence	(5) Interest	(6) Competence
			Panel A: Unweighted regressions			
State fixed effects with covariates						
$T_s \times NCLB_t$	0.046	0.057^{+}	0.010	−0.026	0.026	0.048
	(0.036)	(0.031)	(0.034)	(0.038)	(0.031)	(0.049)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.044	0.055^{+}	0.007	−0.026	0.025	0.048
	(0.035)	(0.032)	(0.036)	(0.037)	(0.031)	(0.049)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.048	0.062	0.019	−0.020	0.026	0.051
	(0.049)	(0.043)	(0.048)	(0.050)	(0.041)	(0.067)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.047	0.061	0.015	−0.020	0.025	0.049
	(0.050)	(0.044)	(0.050)	(0.050)	(0.042)	(0.069)
			Panel B: Weighted regressions			
State fixed effects with covariates						
$T_s \times NCLB_t$	0.065	0.070^{+}	0.015	0.029	0.027	0.045
	(0.046)	(0.039)	(0.055)	(0.066)	(0.046)	(0.056)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.066	0.068^{+}	0.013	0.031	0.028	0.043
	(0.047)	(0.039)	(0.056)	(0.066)	(0.047)	(0.057)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.060	0.057	0.016	0.026	0.021	0.029
	(0.071)	(0.060)	(0.080)	(0.085)	(0.066)	(0.079)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.063	0.061	0.015	0.033	0.024	0.030
	(0.068)	(0.058)	(0.078)	(0.088)	(0.065)	(0.081)
Observations	15,890	15,890	15,890	15,890	15,890	15,890

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{+}p < .10$.

less advantaged students. However, because we conducted multiple hypothesis tests, statistical significance may have occurred by chance. With respect to other findings in the literature, the positive point estimates for engagement are in line with higher teacher-rated engagement due to NCLB from Dee et al. (2013), although Reback et al. (2014) did not find significant effects of NCLB test pressure for enjoyment of math and reading.

While our point estimates are robust across models with and without covariates and across models with either state fixed effects or child fixed effects, there are still some potential concerns that might invalidate the results. First, our main regression results do not account for the fact that four comparison states (Connecticut, Kansas, Michigan, and New York) did not have standardized tests of third graders in the year 2001–2002. Consequently, third graders in those states had not experienced high-stakes testing at the time of the third-grade ECLS-K survey, even though other students in higher grades were taking these tests and their teachers were likely preparing them for tests in the years to come. Second, five treatment states and six comparison states did not test students in fifth grade in 2003–2004. All of these states tested students in fourth grade, so all of the students taking the ECLS-K in fifth grade had experienced taking high-stakes tests the year before. In both these cases, results might be attenuated by including students who were not facing

TABLE 5
Subgroups: Externalizing and Internalizing Results

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
	Socioeconomic status subgroups			
Bottom half of distribution (n = 7,920)				
$T_s \times NCLB_t$	0.023	0.024	−0.017	0.084
	(0.045)	(0.068)	(0.049)	(0.057)
Top half of distribution (n = 7,970)				
$T_s \times NCLB_t$	−0.005	−0.029	−0.041	0.071*
	(0.028)	(0.029)	(0.033)	(0.033)
	Gender subgroups			
Male (n = 7,970)				
$T_s \times NCLB_t$	0.037	−0.001	−0.028	0.089+
	(0.030)	(0.053)	(0.037)	(0.046)
Female (n = 7,920)				
$T_s \times NCLB_t$	−0.020	−0.007	−0.031	0.062
	(0.034)	(0.042)	(0.041)	(0.049)

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. All regression estimates are not weighted and do not include covariates. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
+$p < .10$. *$p < .05$.

TABLE 6
Subgroups: Subject Interest and Competence Results

	Math		Reading		School	
	(1) Interest	(2) Competence	(3) Interest	(4) Competence	(5) Interest	(6) Competence
	Socioeconomic status subgroups					
Bottom half of distribution (n = 7,920)						
$T_s \times NCLB_t$	0.093*	0.086*	0.016	−0.015	−0.002	0.039
	(0.043)	(0.037)	(0.042)	(0.056)	(0.052)	(0.066)
Top half of distribution (n = 7,970)						
$T_s \times NCLB_t$	−0.001	0.027	−0.003	−0.038	0.056	0.060
	(0.043)	(0.036)	(0.042)	(0.034)	(0.044)	(0.048)
	Gender subgroups					
Male (n = 7,970)						
$T_s \times NCLB_t$	0.067+	0.042	0.003	−0.051	0.024	0.060
	(0.036)	(0.042)	(0.034)	(0.041)	(0.039)	(0.052)
Female (n = 7,920)						
$T_s \times NCLB_t$	0.023	0.071+	0.011	−0.001	0.027	0.037
	(0.055)	(0.043)	(0.047)	(0.049)	(0.052)	(0.061)

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. All regression estimates are not weighted and do not include covariates. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
+$p < .10$. *$p < .05$.

pressure of taking a high-stakes test in the contemporaneous year. As a robustness check, we run analyses that exclude the comparison states that did not test third graders in 2001–2002 and states that did not test fifth graders in 2003–2004, and we find that results are very similar to the results when we include all states (see Appendix C).

Limitations

One limitation of the analysis is a lack of precision in the estimates, especially when estimated effect sizes are small. Clustering at the state level produces relatively large standard errors; therefore, one cannot rule out small effects. In models with covariates that do not yield significant effects, we are 95% confident that increased accountability pressure did not affect the outcomes by more than .10 to .17 standard deviations (the range is due to different confidence intervals across different outcomes). The inclusion of child-level fixed effects also increases the size of standard errors; however, the stability of the point estimates across models with state fixed effects and child fixed effects mitigates concerns about bias caused by omitted time-invariant covariates.

As previously stated, although some results are statistically significant at conventional levels, they would lose significance if we accounted for multiple-hypothesis testing. Applying a Bonferroni correction lowers the *p* values needed to attain statistical significance at conventional levels; therefore, our results would no longer be significant.

In addition, as we described in the Methodology section, estimates may be attenuated in this analysis. Moreover, the "true" effect among the treatment states might be smaller than what we would expect for a state with "average" enthusiasm for consequential accountability systems.

The analysis is limited by the socioemotional measures available in the ECLS-K. It would be better to have a data set with a greater range of socioemotional measures. It would also be preferable to have some more "objective" measures of children's changes in behavior and attention, such as psychological tests or classroom observation checklists.

The study only measures the effects of NCLB in the early years of the policy. Consequences for schools became more serious over time; thus, schools might have changed their behavior and affected students more strongly in later years of the policy. Therefore, this study may fail to pick up effects on socioemotional outcomes that might have occurred in later years of the policy.

The subgroup analysis is exploratory and should not be interpreted causally. ECLS-K may lack sufficient power to detect small to moderate effects among subgroups. In addition, the sample weights were not designed to be representative of subgroups, so we run these analyses without sample weights. Furthermore, since NCLB may have affected some comparison states by introducing more stringent subgroup-specific consequences than what those states had prior to NCLB, estimates of subgroup effects may be attenuated.

Implications

The findings from this article suggest several implications for future research. First, it is important for academic researchers to continue studying socioemotional outcomes and noncognitive measures given their impact on academic and adult outcomes (Farrington et al., 2012; Heckman & Kautz, 2012; D. E. Jones, Greenberg, & Crowley, 2015). As national data sets continue to expand the types of socioemotional and noncognitive measures that they collect on surveys, researchers will be able to leverage these measures to complement the extant knowledge about academic outcomes.

Second, it is important that those who collect data on socioemotional and noncognitive outcomes ensure that measures are consistent over time. Although the ECLS-K provides an opportunity to examine socioemotional outcomes between the third and fifth grade, the format of the questionnaire and the wording of questions from the SDQ changed when students were in the eighth grade; consequently, our subscales could not be used to track longer-term impacts of the consequential accountability associated with NCLB.

Third, it is important to continue studying the impact of accountability on socioemotional outcomes. In the extant literature, it is difficult to compare results across studies because of differences in the measures, methodologies, and identification strategies. For example, while Reback et al. (2014) found that increased pressure on students decreased test anxiety (statistically significant), our study finds that academic anxiety increases (not statistically significant in the light of multiple-hypothesis testing). We believe that this discrepancy arises because Reback et al. use a single dichotomized item to measure academic anxiety, whereas we use four Likert-scale items to construct our subscale (see Appendix A)—although if we use a single item to measure test anxiety, our results remain qualitatively similar. In addition, we examine the introduction of NLCB consequential accountability, while Reback et al. examine NCLB pressure.

Finally, future research should examine the effects of consequential accountability on the socioemotional outcomes of children in middle school and high school, as they might be affected very differently by standardized tests than elementary-age children. For example, consequential accountability might have a stronger effect on test anxiety among adolescents, in part due to more personal consequences of scoring poorly on a test, such as lower track placement or, in some cases, failure to graduate from high school. In addition to obvious developmental differences that would affect behavior, younger children tend to have more interest and self-competence in school than older children.

Overall, the findings in this article have potentially important implications for policy and practice. As states create new testing regimes under ESSA, this study suggests that if their new accountability regimes resemble consequential accountability under NCLB, they may not have a strong effect on the type of socioemotional outcomes that we examined for the average student. Thus, states may not need to spend as much time trying to minimize potential negative

consequences of testing on these socioemotional outcomes, although they might make sure to pay closer attention to schools and subgroups that they expect could be more strongly affected. States will need to track how any substantial changes in their policies affect these and other socioemotional outcomes not sufficiently measured on the ECLS-K, such as test anxiety, self-control, and ability to work collaboratively with others.

Appendix A

Outcome Variables: Item Descriptions and Factor Analysis Details

Subscales overview. Each socioemotional outcome variable in this study is a constructed subscale based on survey responses from the ECLS-K SDQ items, which were collected for participating students in Grade 3 in academic year 2001–2002 (i.e., before NCLB took effect) and grade 5 in academic year 2003–2004 (i.e., after NCLB took effect). These data and all item questions are publicly available via the National Center for Education Statistics.[4] Given that the public-use data files contain child-level identifiers, we can merge our constructed subscales with the restricted-use data.

To construct our subscales, we leveraged the publicly available item-level SDQ data and performed exploratory factor analyses on the items that underlie five SDQ composite scales that appear in ECLS-K: Externalizing Problem Behaviors, Internalizing Problem Behaviors, Perceived Interest/Competence in Reading, Perceived Interest/Competence in Math, and Perceived Interest/Competence in All School Subjects.[5] The goal of the exploratory factor analysis was to see if the items in the composite scales were in fact based on the same latent factor. Ultimately, we found that items of these scales better loaded onto separate factors.

Subscale items. We list the SDQ items for each subscale that we constructed. Each subscale is computed as the mean of the item-level responses. As noted in the SDQ documentation files, the SDQ response scale is based on 4 points (1 = *not at all true*, 4 = *very true*).

Externalizing Attention subscale
1. It's hard for me to pay attention.
2. I get distracted easily.
3. It's hard for me to finish my schoolwork.

Externalizing Behavior subscale
1. I get in trouble for talking and disturbing others.
2. I get in trouble for fighting with other kids.

Internalizing Sad/Lonely subscale
1. I often feel lonely.
2. I feel sad a lot of the time.

Internalizing Academic Anxiety subscale
1. I worry about taking tests.
2. I worry about doing well in school.
3. I worry about finishing my work.
4. I feel ashamed when I make mistakes at school.

Perceived Interest in Math subscale
1. I cannot wait to do math each day.
2. I am interested in math.
3. I like math.
4. I enjoy doing work in math.

Perceived Competence in Math subscale
1. Work in math is easy for me.
2. I get good grades in math.
3. I can do very difficult problems in math.
4. I am good at math.

Perceived Interest in Reading subscale
1. I like reading.
2. I am interested in reading.
3. I cannot wait to read each day.
4. I like reading long chapter books.
5. I enjoy doing work in reading.

Perceived Competence in Reading subscale
1. I get good grades in reading.
2. Work in reading is easy for me.
3. I am good at reading.

Perceived Interest in All School Subjects subscale
1. I enjoy work in all school subjects.
2. I like all school subjects.
3. I look forward to all school subjects.

Perceived Competence in All School Subjects subscale
1. I am good at all school subjects.
2. Work in all school subjects is easy for me.
3. I get good grades in all school subjects.

Factor analysis technical details. We conducted a separate exploratory factor analysis on each group of items composing the five composite scales in the SDQ of the ECLS-K. For example, we first ran the analysis on the items in the Externalizing Problem Behaviors scale. Then we ran the analyses on the items in the Internalizing Problem Behaviors scale and on each of the remaining interest/competence scales.

When performing the factor analyses, we used the data only from the third grade. This is an important and purposeful decision: we did not want the underlying latent factors to be contaminated by any potential responses to NCLB. In addition, we used all available data—not just the children in our analytic sample—in conjunction with a third-grade cross-sectional weight because we wanted our generated

TABLE A1
Reliability: Cronbach's Alpha

	Unweighted		Weighted	
Scale: Subscale	Grade 3	Grade 5	Grade 3	Grade 5
Externalizing				
Attention	0.64	0.71	0.63	0.71
Behavior	0.69	0.67	0.69	0.71
Internalizing				
Academic Anxiety	0.73	0.73	0.73	0.72
Sad and Lonely	0.66	0.71	0.65	0.70
Math				
Competence	0.79	0.86	0.78	0.86
Interest	0.90	0.92	0.90	0.92
Reading				
Competence	0.74	0.81	0.74	0.81
Interest	0.84	0.90	0.84	0.89
School				
Competence	0.71	0.78	0.70	0.77
Interest	0.79	0.83	0.78	0.83

subscales to reflect latent factors that are present in the national distribution.

In terms of technical specifications, we first obtained the polychoric correlation matrix of the items for a given subscale. This was a necessary step because the item-level data are on an ordinal scale. After obtaining the polychoric correlation matrix, we performed factor analysis using the matrix. We ran each factor analysis three times, allowing for one, two, and three factors. Ultimately, we found that allowing two factors produced sensible results. Each factor analysis uses a promax rotation of the factors, which allows factors to be correlated.

All items in each scale have factor loadings >0.4 onto a single factor: combined with other information on psychometric properties of the subscales, these subscales appear to be valid and reliable. The subscales were created when multiple items loaded (at least 0.4) onto different factors in third grade.

In Table A1, we provide estimates of the reliabilities for our subscales. The reliabilities range from from 0.62 to 0.92, which are in the acceptable to strong range for research purposes.

Appendix B

Multiple-Hypothesis Testing: Simulation Exercise

In Table 3, we find that coefficient on the $T_s \times NCLB_t$ treatment indicator is statistically significant when academic anxiety is the outcome variable; see column 4 and the regression models that include state fixed effects, with and without covariates. Because conducting multiple hypothesis tests increases the probability of Type I error, we assess whether the observed effect is due to chance. As suggested by an anonymous referee, we run simulations where we randomly assign counterfactual values for T_s, the treatment indicator, to the states in our estimation sample. For each simulation, we stop assigning treatment to additional states once the percentage of students assigned to treatment exceeds the actual percentage of treated students. We then estimate all the regression models for each of the 10 outcomes that we discuss: (1) externalizing: attention, (2) externalizing: behavior, (3) internalizing: academic anxiety, (4) internalizing: sad and lonely, (5) math: competence, (6) math: interest, (7) reading: competence, (8) reading: interest, (9) school: competence, (10) school: interest. In total, we run 5,000 simulations.

TABLE B1
Results of the Simulation Exercise

	Survey sampling weight	
Model type	No	Yes
State fixed effects		
With covariates	0.43	0.61
No covariates	0.48	0.65
Child fixed effects		
With covariates	0.46	0.51
No covariates	0.46	0.58

Note. Percentage of cases in which the lowest p value among any of the 10 estimated treatment effects predicting 10 outcomes is less than the corresponding p value for academic anxiety.

For each simulation, we compute p values for each of the treatment indicators predicting the 10 outcomes and record the lowest p value among the set of indicators; we do this separately for each regression model: state/child fixed effects, with/without covariates, unweighted/weighted. Thus, for each simulation, we are estimating $10 \times 8 = 80$ regressions. After running all simulations, we compute, by regression model, the percentage of times in which the lowest p value from a given simulation is less than the corresponding p value from the academic anxiety models that we estimate in Table 3. A low percentage suggests strong evidence of an effect on academic anxiety, while a high percentage suggests no consistent evidence of an effect.

We report the results of the simulation exercise in Table B1. The percentages are quite high, ranging from 43% to 65%. Therefore, we find that NCLB did not consistently affect academic anxiety.

Appendix C

Additional Robustness Checks

In Tables C1 and C2, we include only states that tested students in fifth grade during the fifth-grade year (2003–2004).

While NCLB accountability was in effect during this year, states were not required to have accountability tests in all relevant grades until 2005–2006 (note that all states in this analysis tested students in fourth grade in 2003–2004). We eliminate 5 treatment states from the model (ME, MO, NJ, OH, WA) and 6 comparison states (CT, MA, MI, NY, RI, WI). This sample allows us to examine the effects NCLB on socioemotional outcomes purely among students being tested in the current year. We expect estimates to either grow or stay the same, because the students left in the sample in the treatment group are experiencing more proximate effects of high-stakes testing, but the same is true of the comparison students left in the sample. As shown in Tables C1 and C2, we find that estimates are generally the same as those that appear in our main tables.

In Tables C3 and C4, we exclude four comparison states that did not have testing in third grade (CT, KS, MI, and NY), since these students may not have experienced as strong a "treatment" prior to NCLB as in the other comparison states. We therefore expect estimates to either grow or stay the same. We find that estimates are qualitatively similar to what appears in our main tables.

In Tables C5 and C6, we switch four states that might have been misclassified as comparison states to treatment states (KS, IN, VA, WI), following a robustness check per Dee and Jacob (2011). It would be problematic if the estimates were to change greatly from the initial model, since we would not want the estimates to be driven by these marginal cases.

TABLE C1
Restricted Sample 1: Externalizing and Internalizing Results

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
	Panel A: Unweighted regressions			
State fixed effects with covariates				
$T_s \times NCLB_t$	0.004	−0.009	−0.010	0.085^{+}
	(0.026)	(0.047)	(0.046)	(0.047)
State fixed effects, no covariates				
$T_s \times NCLB_t$	−0.001	−0.014	−0.013	0.080^{+}
	(0.026)	(0.049)	(0.046)	(0.046)
Child fixed effects with covariates				
$T_s \times NCLB_t$	0.003	−0.003	−0.006	0.083
	(0.036)	(0.069)	(0.062)	(0.067)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	−0.000	−0.004	−0.005	0.085
	(0.037)	(0.068)	(0.062)	(0.066)
	Panel B: Weighted regressions			
State fixed effects with covariates				
$T_s \times NCLB_t$	−0.056	−0.113^{+}	−0.038	0.085
	(0.047)	(0.065)	(0.062)	(0.071)
State fixed effects, no covariates				
$T_s \times NCLB_t$	−0.067	−0.126^{+}	−0.044	0.077
	(0.050)	(0.070)	(0.065)	(0.070)
Child fixed effects with covariates				
$T_s \times NCLB_t$	−0.044	−0.089	−0.024	0.100
	(0.069)	(0.088)	(0.085)	(0.095)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	−0.051	−0.089	−0.021	0.100
	(0.069)	(0.086)	(0.086)	(0.098)
Observations	11,550	11,550	11,550	11,550

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{+}p < .10$.

TABLE C2

Restricted Sample 1: Subject Interest and Competence Results

	Math		Reading		School	
	(1) Interest	(2) Competence	(3) Interest	(4) Competence	(5) Interest	(6) Competence
			Panel A: Unweighted regressions			
State fixed effects with covariates						
$T_s \times NCLB_t$	0.091*	0.094*	−0.005	−0.039	0.062	0.070
	(0.038)	(0.038)	(0.043)	(0.055)	(0.038)	(0.067)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.089*	0.094*	−0.001	−0.035	0.062	0.073
	(0.037)	(0.037)	(0.044)	(0.054)	(0.038)	(0.066)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.090+	0.094+	0.003	−0.036	0.057	0.066
	(0.052)	(0.054)	(0.057)	(0.072)	(0.052)	(0.092)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.091+	0.097+	0.007	−0.033	0.059	0.070
	(0.052)	(0.053)	(0.058)	(0.073)	(0.052)	(0.093)
			Panel B: Weighted regressions			
State fixed effects with covariates						
$T_s \times NCLB_t$	0.053	0.038	−0.048	−0.024	0.048	0.066
	(0.054)	(0.048)	(0.068)	(0.096)	(0.058)	(0.085)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.053	0.042	−0.044	−0.016	0.045	0.070
	(0.054)	(0.048)	(0.069)	(0.093)	(0.060)	(0.084)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.040	0.015	−0.062	−0.039	0.024	0.034
	(0.081)	(0.075)	(0.088)	(0.121)	(0.083)	(0.123)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.046	0.025	−0.055	−0.033	0.033	0.042
	(0.080)	(0.073)	(0.086)	(0.122)	(0.081)	(0.125)
Observations	11,550	11,550	11,550	11,550	11,550	11,550

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.

+$p < .10$. *$p < .05$.

TABLE C3

Restricted Sample 2: Externalizing and Internalizing Results

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
		Panel A: Unweighted regressions		
State fixed effects with covariates				
$T_s \times NCLB_t$	0.018	0.003	−0.027	0.089*
	(0.024)	(0.039)	(0.033)	(0.040)
State fixed effects, no covariates				
$T_s \times NCLB_t$	0.017	0.001	−0.027	0.086*
	(0.025)	(0.041)	(0.033)	(0.039)

(continued)

TABLE C3 (CONTINUED)

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
Child fixed effects with covariates				
$T_s \times NCLB_t$	0.022	0.008	−0.020	0.088
	(0.034)	(0.056)	(0.043)	(0.055)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	0.016	0.009	−0.023	0.089
	(0.034)	(0.056)	(0.044)	(0.056)
	Panel B: Weighted regressions			
State fixed effects with covariates				
$T_s \times NCLB_t$	−0.001	−0.043	−0.049	0.146^{+}
	(0.051)	(0.064)	(0.059)	(0.079)
State fixed effects, no covariates				
$T_s \times NCLB_t$	−0.006	−0.053	−0.052	0.141^{+}
	(0.052)	(0.065)	(0.060)	(0.078)
Child fixed effects with covariates				
$T_s \times NCLB_t$	0.005	−0.019	−0.038	0.154
	(0.071)	(0.084)	(0.081)	(0.106)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	−0.003	−0.025	−0.043	0.151
	(0.071)	(0.083)	(0.083)	(0.110)
Observations	14,110	14,110	14,110	14,110

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{+}p < .10$. $^{*}p < .05$.

TABLE C4
Restricted Sample 2: Subject Interest and Competence Results

	Math		Reading		School	
	(1) Interest	(2) Competence	(3) Interest	(4) Competence	(5) Interest	(6) Competence
	Panel A: Unweighted regressions					
State fixed effects with covariates						
$T_s \times NCLB_t$	0.065^{+}	0.081**	0.009	−0.025	0.033	0.067
	(0.035)	(0.029)	(0.036)	(0.041)	(0.033)	(0.049)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.064^{+}	0.079**	0.005	−0.026	0.031	0.065
	(0.035)	(0.030)	(0.038)	(0.040)	(0.033)	(0.050)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.067	0.085*	0.013	−0.023	0.030	0.067
	(0.048)	(0.041)	(0.050)	(0.055)	(0.044)	(0.068)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.067	0.084*	0.011	−0.022	0.030	0.064
	(0.048)	(0.042)	(0.052)	(0.055)	(0.046)	(0.070)

(continued)

TABLE C4 (CONTINUED)

	Math		Reading		School	
	(1) Interest	(2) Competence	(3) Interest	(4) Competence	(5) Interest	(6) Competence
Panel B: Weighted regressions						
State fixed effects with covariates						
$T_s \times NCLB_t$	0.061	0.079^{+}	0.025	0.026	0.039	0.058
	(0.049)	(0.040)	(0.056)	(0.068)	(0.049)	(0.057)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.063	0.077^{+}	0.022	0.026	0.037	0.052
	(0.050)	(0.041)	(0.057)	(0.067)	(0.050)	(0.059)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.053	0.064	0.019	0.017	0.028	0.038
	(0.074)	(0.063)	(0.080)	(0.087)	(0.070)	(0.081)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.061	0.070	0.023	0.029	0.033	0.039
	(0.072)	(0.061)	(0.080)	(0.091)	(0.069)	(0.083)
Observations	14,110	14,110	14,110	14,110	14,110	14,110

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

TABLE C5
Reclassifying States: Externalizing and Internalizing Results

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
Panel A: Unweighted regressions				
State fixed effects with covariates				
$T_s \times NCLB_t$	0.039^{+}	0.011	−0.012	0.069*
	(0.023)	(0.035)	(0.030)	(0.035)
State fixed effects, no covariates				
$T_s \times NCLB_t$	0.038	0.010	−0.014	0.068^{+}
	(0.024)	(0.037)	(0.030)	(0.035)
Child fixed effects with covariates				
$T_s \times NCLB_t$	0.038	0.014	−0.009	0.066
	(0.032)	(0.050)	(0.040)	(0.048)
Child fixed effects, no covariates				
$T_s \times NCLB_t$	0.036	0.015	−0.011	0.066
	(0.033)	(0.049)	(0.040)	(0.050)
Panel B: Weighted regressions				
State fixed effects with covariates				
$T_s \times NCLB_t$	0.025	−0.060	−0.010	0.085
	(0.046)	(0.055)	(0.052)	(0.068)
State fixed effects, no covariates				
$T_s \times NCLB_t$	0.019	−0.068	−0.014	0.080
	(0.047)	(0.057)	(0.053)	(0.067)
Child fixed effects with covariates				
$T_s \times NCLB_t$	0.024	−0.052	−0.003	0.091
	(0.064)	(0.071)	(0.072)	(0.094)

(continued)

TABLE C5 (CONTINUED)

	Externalizing		Internalizing	
	(1) Attention	(2) Behavior	(3) Sad/lonely	(4) Academic anxiety
Child fixed effects, no covariates				
$T_s \times NCLB_t$	0.019	−0.055	−0.004	0.088
	(0.065)	(0.071)	(0.074)	(0.096)
Observations	15,890	15,890	15,890	15,890

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{+}p < .10$. $^{*}p < .05$.

TABLE C6
Reclassifying States: Subject Interest and Competence Results

	Math		Reading		School	
	(1) Interest	(2) Competence	(3) Interest	(4) Competence	(5) Interest	(6) Competence
Panel A: Unweighted regressions						
State fixed effects with covariates						
$T_s \times NCLB_t$	0.029	0.064*	0.008	−0.037	0.018	0.052
	(0.035)	(0.031)	(0.032)	(0.036)	(0.031)	(0.044)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.029	0.064*	0.006	−0.038	0.017	0.051
	(0.035)	(0.031)	(0.033)	(0.036)	(0.030)	(0.044)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.029	0.067	0.012	−0.033	0.015	0.052
	(0.049)	(0.044)	(0.045)	(0.049)	(0.041)	(0.061)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.029	0.069	0.012	−0.032	0.016	0.052
	(0.050)	(0.044)	(0.047)	(0.049)	(0.042)	(0.062)
Panel B: Weighted regressions						
State fixed effects with covariates						
$T_s \times NCLB_t$	0.060	0.079*	0.001	0.026	0.001	0.075
	(0.048)	(0.037)	(0.051)	(0.059)	(0.050)	(0.051)
State fixed effects, no covariates						
$T_s \times NCLB_t$	0.060	0.078*	0.001	0.028	0.001	0.075
	(0.048)	(0.038)	(0.051)	(0.060)	(0.050)	(0.051)
Child fixed effects with covariates						
$T_s \times NCLB_t$	0.048	0.066	−0.002	0.022	−0.004	0.067
	(0.073)	(0.058)	(0.074)	(0.079)	(0.070)	(0.071)
Child fixed effects, no covariates						
$T_s \times NCLB_t$	0.052	0.071	−0.000	0.029	0.000	0.071
	(0.070)	(0.056)	(0.072)	(0.081)	(0.069)	(0.073)
Observations	15,890	15,890	15,890	15,890	15,890	15,890

Note. Number of observations is rounded to the nearest 10 to comply with National Center for Education Statistics restricted-use data reporting standards. Weighted regressions are based on a third- through fifth-grade longitudinal weight to achieve national representation. Time-invariant covariates include student-level controls for race/ethnicity, socioeconomic status, and first-grade *t* scores in math and reading. Time-varying covariates include school enrollment, proportion Hispanic and Black, proportion of students who are eligible for free and reduced-price lunch, an indicator for whether the school is a charter school, and an indicator for whether the child switched schools between third and fifth grade. All models include cluster-robust standard errors in parentheses; clustering is at the state level.
$^{*}p < .05$.

Authors' Note

Both authors contributed equally to this article. All errors are our own.

Notes

1. Academic year 2001–2002 was the first year of required testing to establish a baseline for scores at each school against which progress would be measured in subsequent years. However, testing was not required at each of Grades 3 to 8 until 2005–2006, and there was variation across states in which grades were tested in each year. We provide detail on this variation, and we examine implications for this article's analysis and for interpretation in the Discussion section.

2. In some states, only schools receiving Title I aid were subject to sanctions, while in other states any school was potentially subject to sanctions (Dee & Jacob, 2011).

3. The authors do not specify which items they selected from the survey to create the enjoyment scales, because they were prevented from doing so by copyright rules.

4. ECLS-K SDQ data website: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010070.

5. Specific items that belong to each of these composite scales is publicly available via the National Center for Education Statistics: http://nces.ed.gov/pubs2010/data/2010070_sdq_readme.pdf.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors' work on this study was funded in part by the U.S. Department of Education, Institute of Education Sciences (Grant #R305B090016).

References

Atkins-Burnett, S., & Meisels, S. J. (2001). *Measures of socio-emotional development in middle childhood* (Working Paper No. 2001-03). Washington, DC: National Center for Education Statistics.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*(1), 249–275.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The impact of assessment and accountability on teacher recruitment and retention are there unintended consequences? *Public Finance Review*, *36*(1), 88–111.

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

Dee, T. S., & Jacob, B. A. (2010). The impact of No Child Left Behind on students, teachers, and schools. *Brookings Papers on Economic Activity*, *2010*(2), 149–194.

Dee, T. S., & Jacob, B. A. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446.

Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252–279.

Downer, J. T., Rimm-Kaufman, S. E., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's behavioral engagement in learning? *School Psychology Review*, *36*(3), 413.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review*. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Feng, L., Figlio, D. N., & Sass, T. (2010). *School accountability and teacher mobility* (Working Paper No. 16070). Cambridge, MA: National Bureau of Economic Research.

Hannaway, J., & Hamilton, L. (2008). *Performance-based accountability policies: Implications for school and classroom practices*. Washington, DC: Urban Institute and RAND Corporation.

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

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, *118*(3), 843–877.

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health*, *105*(11), 2283–2290.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan*, *81*(3), 199–203.

Kovacs, M., & Devlin, B. (1998). Internalizing disorders in childhood. *Journal of Child Psychology and Psychiatry*, *39*(1), 47–63.

Leadbeater, B. J., Blatt, S. J., & Quinlan, D. M. (1995). Gender-linked vulnerabilities to depressive symptoms, stress, and problem behaviors in adolescents. *Journal of Research on Adolescence*, *5*(1), 1–29.

Li, D. (2012). *School accountability and principal mobility: How No Child Left Behind affects the allocation of school leaders*. Retrieved from https://sites.google.com/site/danielleli/research

Madaus, G. F. (1991). The effects of important tests on students: Implications for a national examination system. *Phi Delta Kappan*, *73*(3), 226–231.

McMurrer, J. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Washington, DC: Center on Education Policy.

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

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, *20*(5), 12–20.

Polikoff, M. S., & Struthers, K. (2013). Changes in the cognitive complexity of English instruction: The moderating effects of school and classroom characteristics. *Teachers College Record*, *115*(8).

Pollack, J. M., Atkins-Burnett, S., Najarian, M., & Rock, D. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) psychometric report for the fifth grade*. (Technical/Methodological Report No. NCES 2006036REV). Washington, DC: National Center for Education Statistics.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5), 1394–1415.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, *6*(3), 207–241.

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist*, *40*(1), 53–63.

Segool, N. K., Carlson, J. S., Goforth, A. N., Von Der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489–499.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research*, *4*(1), 27–41.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in experimental social psychology*, *21*, 261–302.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242.

Wong, M., Cook, T. D., & Steiner, P. M. (2009). *No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series* (Working Paper No. WP-09-11). Evanston, IL: Institute for Policy Research.

Wong, M., Cook, T. D., & Steiner, P. M. (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.

Authors

CAMILLE R. WHITNEY is an independent educational researcher. Her research focuses on mindfulness for educators and students, as well as student social and emotional outcomes and attendance.

CHRISTOPHER A. CANDELARIA is an assistant professor of public policy and education at Vanderbilt University. His research interests include school finance, teacher labor markets, and accountability.