American Educational Research Journal
June 2022, Vol. 59, No. 3, pp. 610–647
DOI: 10.3102/00028312211058460
Article reuse guidelines: sagepub.com/journals-permissions
© 2021 AERA. https://journals.sagepub.com/home/aer

Did States' Adoption of More Rigorous Standards Lead to Improved Student Achievement? Evidence From a Comparative Interrupted Time Series Study of Standards-Based Reform

Mengli Song
Michael S. Garet
American Institutes for Research
Rui Yang
Facebook
Drew Atchison
American Institutes for Research

This study was designed to assess the effects of states' adoption of more rigorous standards as part of the current wave of standards-based reform on student achievement using comparative interrupted time series analyses based on state-level NAEP data from 1990 to 2017. Results show that the effects of adopting more rigorous standards on students' mathematics achievement were generally small and not significant. The effects on students' reading achievement were also generally small, but negative and statistically significant for Grade 4. The study also revealed that the effects of states' adoption of more rigorous standards varied across NAEP subscales and student subgroups.

Mengli Song, PhD, is a principal researcher at the American Institutes for Research; email: *msong@air.org*. Her work focuses on research design, quantitative methods, and evaluations of educational programs and policy.

Michael S. Garet, PhD, is a vice president and Institute Fellow at the American Institutes for Research. He has designed and led rigorous studies of teacher professional development and teacher and leader evaluation systems, as well as continuous improvement in schools.

Rui Yang, PhD, is a researcher at Facebook. His work focuses on applying quantitative methods to solving people and business problems.

Drew Atchison, PhD, is a senior researcher at the American Institutes for Research. In that role, he conducts quantitative analysis on a wide range of topics, including school and district improvement, accountability, education finance, and educational equity.

Keywords: college- and career-ready standards, comparative interrupted time series design, student achievement, NAEP

Introduction

Since the release of the report *A Nation at Risk* (National Commission on Excellence in Education, 1983), standards-based reforms have been a crucial part of federal and state efforts to improve education. An important impetus for the current wave of standards-based reform was the recognition that the rigor of states' standards varied widely across states and declined in many states as an unintended consequence of the accountability requirements under the No Child Left Behind Act of 2001 (Bandeira de Mello et al., 2009). To encourage states to adopt more rigorous standards, the Obama administration built into its $4.35 billion Race To the Top grant program the requirement that states applying for the grant need to demonstrate their commitment to adopting rigorous standards. Moreover, states were required to participate in a consortium consisting of a significant number of states working toward jointly developing and adopting a common set of high-quality internationally benchmarked K–12 standards (U.S. Department of Education, 2009).

The push for common standards across states was further strengthened through the Elementary and Secondary Education Act (ESEA) flexibility, which provided states with waivers of certain ESEA requirements. As one condition for receiving the waivers, states were required to adopt so-called "college- and career-ready (CCR) standards," formally defined as "content standards for kindergarten through 12th grade that build toward college and career readiness by the time of high school graduation" (U.S. Department of Education, 2012, p. 5). According to the ESEA flexibility requirement, a state's CCR standards must be either (1) standards that are common to a significant number of states; or (2) standards that are approved by a state network of institutions of higher education, which must certify that students who meet the standards will not need remedial course work at the postsecondary level. In total, 45 states, the District of Columbia (DC), and two U.S. territories submitted requests for ESEA flexibility, and approval was granted to all but two states. The great majority of those states met the flexibility requirement regarding CCR standards by adopting the Common Core State Standards (CCSS).

Indeed, a strong emphasis on common standards across states—the CCSS in particular—has been a distinctive feature of the latest wave of standards-based reform. Released in June 2010, the CCSS was quickly adopted by 45 states and DC by the end of 2011 and adopted by one more state (Washington) in June 2012.[1] The other four states adopted their own CCR standards by 2015. The extraordinary initial response of states to the CCSS, however, was followed by a steady decline in public support (Henderson et al., 2020). By fall 2017, 11 states had announced a major Common Core rewrite or replacement (Education Week, 2017). There are many reasons for the widespread opposition to the CCSS; the lion's share of attacks on the CCSS, however, are grounded in politics (Gewertz, 2015; Polikoff et al., 2016).

While there is no shortage of opinions in the contentious field of standards-based reform, there is clearly a lack of rigorous empirical evidence on the impact of the new standards on student learning. The study presented in this article is intended to begin to fill in this gap. As part of a larger research agenda on the implementation and impact of CCR standards, this study was designed to answer the overarching question–*Did states' adoption of more rigorous standards as part of the current wave of standards-based reform result in increases in student achievement in reading and mathematics, both overall and for key student subgroups?* Relying on 1990–2017 state-level NAEP data in reading and mathematics for Grades 4 and 8, we addressed this question using a comparative interrupted time series (CITS) design. Before we describe in detail the study design and present study findings, a brief review of relevant research is in order.

Review of Relevant Research

During the past few years, there has been a growing body of research on the current wave of standards-based reform, which was grounded in statewide adoption of rigorous CCR standards. Given that the CCSS has been the dominant form of CCR standards, and given its highly contentious nature, it is no surprise that existing research in this area has focused almost exclusively on the CCSS rather than other types of CCR standards. Most of the studies on the CCSS examined the implementation of the standards, with only a few designed to assess the impact of the standards on teaching and learning. In this section, we first briefly review findings from existing research on the implementation of the CCSS, which provides useful context for understanding the findings from this study. We then review research on the impact of the CCSS based on both teachers' self-report and analyses of student achievement data.

Research on the Implementation of the CCSS

Most of the existing studies of CCSS implementation are survey-based descriptive studies, focusing in particular on implementation challenges. These surveys have revealed that statewide transition from older standards to the new CCR standards has created a multitude of challenges at multiple levels. At the state level, finding adequate resources to support all the necessary CCSS implementation activities was the most frequently cited challenge faced by states based on the annual surveys conducted by the Center on Education Policy (CEP). Among the 40 CCSS-adopting states that responded to the 2013 CEP survey, for example, 22 states considered inadequate funding as a major challenge in implementing the CCSS (Rentner, 2013), a figure similar to that found in earlier CEP surveys (Kober & Rentner, 2011, 2012).

Most of the states participating in the 2013 CEP state survey also reported challenges such as developing educator evaluation systems that hold educators accountable for student mastery of the CCSS (32 states) and identifying and/or developing curriculum materials needed for implementing the new standards (26 states; Rentner, 2013). Furthermore, the majority of the states surveyed reported challenges in providing sufficient professional development on the CCSS for teachers (37 states) and principals (33 states; Kober et al., 2013).

At the district level, finding adequate resources also topped the list of challenges in CCSS implementation. Based on the 2014 CEP district survey (Rentner & Kober, 2014), inadequate resources were cited as a major implementation challenge by two thirds of the districts and a minor challenge by one quarter of the districts in CCSS-adopting states. A large majority of the districts surveyed also identified the following issues as either a major or a minor challenge: having adequate district staffing levels (87%) and staff expertise (86%) to implement the CCSS, identifying and/or developing the curriculum materials necessary to implement the CCSS (90%), providing high-quality professional development and other support to help teachers implement the CCSS instructional activities (88%), and having enough time to implement the CCSS before consequences are tied to student performance on the CCSS-aligned assessment (89%).

Implementation of the CCSS at the school level also proved to be challenging. According to surveys conducted in 2013 and 2014 by Scholastic and Bill & Melinda Gates Foundation, almost three quarters (73%) of teachers surveyed reported in 2013 that they believed the implementation of the CCSS in their schools was or would be challenging, and the sentiment was shared even more widely in 2014 (81%). The teachers responding to the 2014 survey cited a variety of issues that had been problematic for their schools in implementing the CCSS, most notably factoring student results on new tests into teacher evaluation (59%) and uncertainty about which assessments their state will use (51%).

Finally, at the teacher level, one major challenge to successful implementation of the CCSS was that many teachers were not well prepared to implement the new standards. The Scholastic and Bill & Melinda Gates Foundation (2014) survey, for instance, revealed that less than one third (31%) of the teachers surveyed felt "very" prepared to teach the CCSS, less than half (48%) felt only "somewhat" prepared," and 21% felt "somewhat" or "very" unprepared 4 years after the release of the CCSS.

At the teacher/classroom level, another obstacle to successful transition to the CCSS was the lack of CCSS-aligned curricular and instructional materials. In a 2016 survey of K–8 math teachers from 43 CCSS-adopting states and DC, over 40% of the teachers surveyed reported that the available math materials were not well aligned to the new standards (Bay-Williams et al., 2016). Several content analyses of textbooks reached a similar conclusion. The first round of reviews of K–8 math instructional series released by EdReport.org, for instance, showed that, contrary to the publishers' claims, 17 of the 20 math curricula reviewed failed to meet criteria for alignment with the CCSS (Heitin, 2015). Findings from EdReport.org's initial review of seven English language arts (ELA) series were mixed but more positive—three were considered fully aligned to the CCSS, three partially aligned, and one fully unaligned (Heitin, 2016). These findings mirror the findings from Polikoff's (2015) analysis of the alignment of four popular textbooks to the CCSS for Grade 4 mathematics, which revealed areas of substantial misalignment and challenged the publishers' claims of alignment.

Research on the Impact of the CCSS

While there has been a large body of research examining the implementation of the CCSS, research on the impact of the CCSS is rather limited, possibly due to challenges in designing rigorous impact studies given the nearly universal adoption of the CCSS. In this section, we review findings from a number of survey-based studies which gathered data on the impact of the CCSS on teaching and learning as reported by teachers. These "impact" findings are descriptive in nature and thus do not warrant causal conclusions. Also reviewed in this section are findings from four studies that explicitly assessed the impact of the CCSS based on student achievement data (Loveless, 2014, 2015, 2016; Xu & Cepa, 2018). Given design limitations, however, findings from these studies also need to be interpreted with caution. Moreover, findings from these four studies are limited in that they examined the impact of the CCSS only in the years immediately after adoption, even though full implementation of ambitious educational policy typically takes many years to accomplish (Coburn et al., 2016).

Impact of the CCSS on Teaching and Learning as Reported by Teachers

There has been evidence based on teachers' self-reports that CCSS implementation has produced positive changes in both teachers' instructional practice and student learning. More than three quarters (76%) of the teachers participating in the survey conducted by Kane et al. (2016), for example, reported having changed at least half of their classroom instruction as a result of the CCSS. The majority of the K–8 math teachers responding to the survey conducted by Bay-Williams et al. (2016) similarly reported making changes to many of their practices in ways consistent with the CCSS. Almost two thirds (64%) of the teachers surveyed, for example, reported that they were devoting more attention to requiring students to explain in writing how they got their answers than before the CCSS were introduced, and 55% of the teachers reported a greater focus on requiring students to use proper math vocabulary than before.

Teachers also reported positive changes in student learning. Even though the CCSS was not expected to be fully implemented until the 2013–2014 school year in many states, over half (53%) of the teachers in CCSS-adopting states reported in 2013 that they had already seen a positive change in their students' ability to think critically and use reasoning skills as a result of the CCSS (Scholastic & the Bill & Melinda Gates Foundation, 2014). Teachers' views on the impact of the CCSS were most positive among elementary school teachers—62% of elementary teachers reported having seen a positive impact on students' ability to think critically and use reasoning skills, as compared with 47% for middle school teachers and 37% for high school teachers. Among the K–8 math teachers surveyed by Bay-Williams et al. (2016), those teaching lower grades similarly expressed more positive views of the impact of the CCSS on student learning than those teaching higher grades. The authors attributed the finding to the fact that the new middle school standards are much harder than the new elementary school standards, particularly relative to the standards that they replaced.

Impact of the CCSS on Student Achievement

To date, there have been only a handful of studies that attempted to assess the impact of the CCSS on student achievement (Loveless, 2014, 2015, 2016; Xu & Cepa, 2018). As part of the annual Brown Center reports on American education, the three studies conducted by Loveless were intended to estimate the CCSS's early impact by comparing changes in NAEP test scores between states with different levels of CCSS implementation. Relying on a measure of "congruence" or similarity between each state's 2009 mathematics standards and the CCSS for mathematics created by Schmidt and Houang (2012), Loveless (2014) compared the 2009–2013 eighth-grade NAEP gains across five groups of states with congruence ratings ranging from 1 (i.e., "least like CCSS") to 5 (i.e., "most like CCSS"), and found no systematic relationship between states' congruence ratings and changes in their NAEP scores.

In a second set of analyses, Loveless (2014) used a CCSS implementation index created based on a 2011 survey of state education agencies to classify states into three groups: strong implementers (*n* = 19), medium implementers (*n* = 26), and nonadopters (*n* = 5). He concluded that strong implementers experienced a larger improvement in NAEP scores in eighth-grade math from 2009 to 2013 than did nonadopters. The difference, although in the desired direction, was very small (1.27 points, or 0.04 standard deviations [*SD*s], on the NAEP scale).

The same conclusion was also reached in Loveless' (2015) study, in which Loveless replicated his 2014 analyses using NAEP data for fourth-grade reading, and conducted similar analyses using an alternative CCSS implementation index that designated states as strong implementers (n = 12), medium implementers (*n* = 34), and nonadopters (*n* = 4) based on whether the state was expected to fully implement the new standards by the 2012–2013 school year. Both sets of analyses suggest that the 2009–2013 gain in NAEP fourth-grade reading score was only slightly higher (by 0.03–0.04 *SD*s) in strong implementers than in nonadopting states. Similar analyses based on the two alternative CCSS implementation indices and 2009–2015 NAEP data, however, revealed that the 2009–2015 gain in NAEP fourth-grade reading score was actually slightly smaller in strong implementers than in nonadopting states—by 0.01–0.02 *SD*s (Loveless, 2016). For eighth-grade math, the 2009–2015 NAEP gain in strong implementers was slightly smaller (by 0.003 *SD*s) based on one implementation index and slightly larger (by 0.02 *SD*s) based on the other implementation index relative the gain in nonadopting states.

Taken together, the three studies conducted by Loveless suggest that there were very little systematic differences between states that were strong CCSS implementers and nonadopting states in NAEP gains between 2009 and 2015. These findings, however, need to be interpreted with caution, as they were based on simple descriptive comparisons of group means between nonequivalent groups of states without any control for potential selection bias. Thus, those findings reflect associations rather than causal effects. In particular, the "control group" used in all three studies included a small set of nonadopting states, which were quite unique given the almost nationwide adoption of the CCSS. These nonadopters therefore may not be an appropriate control group as selection bias may be a serious concern. In addition, given the very small number (as few as 4) of states in the control group, results from the analyses conducted by Loveless (2014, 2015, 2016) were sensitive to substantial changes in NAEP scores in one or two states, as the author acknowledged.

While the three studies discussed above analyzed NAEP data from all 50 states, the study conducted by Xu and Cepa (2018) focused on the early effect of the CCSS in a single state—Kentucky. In this study, the authors tracked three cohorts of students from Grade 8 to Grade 11 and found that students exposed to the CCSS (i.e., students in the two more recent cohorts) scored significantly higher on the ACT taken in the 11th grade than similar students in the earliest cohort not exposed to the new standards (differences = 0.03~0.04 *SD*s). The authors cautioned, however, that the observed differences between the cohorts may not be completely attributable to the CCSS implementation, as cross-cohort differences in student achievement occurred in both the year before and the year after the adoption of the CCSS.

Clearly, despite the intense interest from both policy makers and the general public in the CCSS, the empirical research base on the impact of the standards on student achievement is still quite thin and does not yet contain convincing evidence on the impact of the CCSS due to various design challenges and study limitations (Polikoff, 2017). Moreover, focusing exclusively on the CCSS, existing research has not yet tested the basic premise underlying the current wave of standards-based reform–that is, adopting more rigorous standards would lead to improved student achievement. The study presented in this article is intended to fill in this gap in the literature by providing empirical evidence on this untested premise, drawing on longitudinal state-wide NAEP data collected up to date and analyzed based on a rigorous quasi-experimental design that aimed to maximize the internal validity of study findings in a research context where a gold-standard randomized controlled trial was not feasible.

Method

In this section, we first present an overview of the study design. We then describe in detail the measures, data sources, and analytic approach that we used to address the overarching research question guiding this study.

Design Overview

To assess whether states' adoption of more rigorous standards as part of the current wave of standards-based reform led to improved student achievement, we analyzed state-level NAEP data between 1990 and 2017 using a CITS design, a quasi-experimental design often used to assess the effects of programs and policies that do not lend themselves to randomized experiments. The CITS design is an extension of the interrupted time series design, which, in its simplest form, measures an outcome for a treatment group multiple times before and after the treatment starts (i.e., the point of "interruption") and then estimate the treatment effect as the deviation in the level and/or slope of the outcome from before to after the onset of the treatment.

The simple interrupted time series design, however, is subject to various threats to internal validity, particularly threats due to history—in this case, the possibility that forces other than the adoption of more rigorous standards might have influenced student achievement after the new standards were introduced (Shadish et al., 2002). To guard against potential threats to internal validity, a comparison group is often added to this simple version of time series design, extending it to a CITS design (Wong et al., 2015). Identifying a plausible comparison group that was not affected by the treatment under study, as Dee and Jacob (2011) noted, is the central challenge for any CITS design. It is particularly challenging to identify an appropriate comparison group for this study, as all 50 states and DC adopted supposedly more rigorous new standards—CCR standards—by 2015.[2]

To assess the effects of states' adoption of more rigorous standards, one obvious approach is to take advantage of the natural variation between states in the timing of CCR standards adoption and compare the achievement trend between states that had adopted the new CCR standards (i.e., treatment states) and states that had not yet adopted the standards (i.e., comparisons states). This approach relies on the assumption that there is sufficient variation between states in the timing of CCR standards adoption, which unfortunately does not hold. Even though states officially adopted CCR standards in ELA over a 6-year window (2008–2014) and CCR standards in mathematics over an 8-year window (2007–2015), the overwhelming majority of states (41 for ELA and 39 for math) adopted CCR standards in the same year—2010 (see the appendix for the year of CCR standards adoption for each state).

Given the limited variation in the timing of CCR standards adoption across states, we used an alternative approach to assessing the effects of the adoption of more rigorous standards on student achievement, taking advantage of the natural variation between states in the quality of their content standards prior to the adoption of the new CCR standards. For this study, we relied on the following two existing measures of the quality of states' prior content standards:

- *Prior rigor index*: A measure of the rigor of each state's 2010 standards created by the Thomas Fordham Institute (Carmichael et al., 2010), with separate ratings for ELA and mathematics standards.[3]
- *Prior CCSS-similarity index*: A measure of the similarity between each state's 2009 mathematics standards and the CCSS for mathematics created by researchers at Michigan State University (Schmidt & Houang, 2012).

Next, we explain how we classified states into treatment states and comparison states for our CITS analyses based on these measures. (See the appendix for how each state's prior standards are rated on each measure and the treatment status of each state determined by each measure.)

State Classification Based on the Prior Rigor Index

The prior rigor index rates the content standards of each state as of May 2010 for their content and rigor based on a detailed scoring rubric developed by the Thomas Fordham Institute, which has been engaging in appraising state academic standards since the late 1990s. The rating is on a 0–7 point scale and measures the extent to which a state's standards address content-specific criteria for a given subject area fully, with high quality, and with the level of rigor appropriate for the target grades (see Appendix A in Carmichael et al. [2010] for the list of content-specific criteria for ELA and mathematics and details of the scoring rubric). For our main CITS analyses based on this index, we defined treatment states as states with a score of 0–3 and comparison states as states with a score of 5–7 on the index. We excluded states with a score of 4 from our main analyses to allow for a sharper treatment contrast and included those states in a set of sensitivity analyses. Our assumption is that CCR standards would represent a stronger form of "treatment" for states with less rigorous prior standards than for states that already had fairly rigorous standards in place before CCR standards were adopted. Thus, we expected that the new standards, if well implemented, would lead to a larger improvement in student achievement in states with less rigorous prior standards than in states with more rigorous prior standards.

State Classification Based on the Prior CCSS-Similarity Index

The prior CCSS-similarity index measures the similarity between a state's 2009 standards and the CCSS in mathematics in terms of the focus and coherence of the topics covered in the standards. The measure was created using well-established methods developed for the analysis of standards documents in the Third International Mathematics and Science Study (TIMSS; Schmidt et al., 1997). Based on this measure, Schmidt and Houang (2012) grouped states into five categories, ranging from "least like CCSS" to "most like CCSS." Given that the CCSS for mathematics has the highest rating on the prior rigor index (see Note 3), it is reasonable to assume that a state's prior math standards were less rigorous if they were less like the CCSS; thus the prior CCSS-similarity index can be viewed as an indirect measure of the rigor of a state's prior standards.[4] For our main CITS analyses, we defined treatment states as states in the two "least like CCSS" groups, and comparison states as states in the two "most like CCSS" groups. We excluded states in the middle group to allow for a sharper contrast and included these states in our sensitivity analyses. Our assumption here is that the new CCR standards reflect a larger increase in the rigor of state standards and thus a stronger form of treatment for states whose prior standards were less like the CCSS than for states whose prior standards were more like the CCSS. Consequently, we expected the new standards to lead to a larger improvement in student achievement in the treatment states than in the comparison states defined based on the prior CCSS-similarity index.

Given that the overwhelming majority of states adopted CCR standards between June and November of 2010, and given that the prior rigor index pertains to states' standards as of May 2010 and the prior CCSS-similarity index pertains to states' 2009 standards, we restricted our CITS analyses to states that adopted CCR standards in 2010. Specifically, our main CITS analyses based on the prior rigor index included 17 treatment states and 12 comparison states for reading and 20 treatment states and 14 comparison states for math. Our main analyses based on the prior CCSS-similarity index, which is available for math only, included 14 treatment states and 12 comparison states.

All states included in the two math CITS analysis samples and all but Virginia included in the reading analysis sample are states that adopted the CCSS,[5] although not all these states have stuck to the CCSS after adoption. By late 2016, of the 38 states included in our main CITS analyses that originally adopted the CCSS, four had officially repealed the CCSS and replaced it with their own version of CCR standards, nine states had completed revisions to the CCSS, 11 states had been undergoing a CCSS revision process or announced their intention to repeal or replace the CCSS.[6] Thus, not all CCSS-adopting states have been implementing the original version of the CCSS verbatim; some have been implementing a revised version of the CCSS or their own version of CCR standards. However, the revisions states made to the CCSS were generally minor and primarily clarifications in nature, according to an in-depth analysis of CCSS revisions in nine states conducted by Norton et al. (2016).

Based on the state classifications described above, we assessed the effects of states' adoption of more rigorous standards on student achievement by comparing treatment states and comparison states in the change in their student achievement trajectories from before to after the 2010 adoption of CCR standards. Given that the NAEP assessments in both reading and math have been administered every other year since 2003, the available NAEP data (1990–2017) allowed us to estimate the effects of adopting more rigorous standards for states included in our CITS analyses 1 year, 3 years, 5 years, and 7 years after the 2010 adoption of the new CCR standards (i.e., 1-/3-/5-/7-year effect).[7] Although we learned that most states expected to take 3 to 5 years to reach full implementation of the new standards (see Supplemental Exhibits 1a and 1b in the online version of the journal), the present study did not track the actual implementation of the standards in each state. Thus, the 1-/3-/5-/7-year effects estimated in this study do not represent the effects of 1/3/5/7 years of full implementation of the new standards.

Data and Measures

Measures of Student Achievement

Our primary measures of student achievement are state average NAEP scores in reading and math for Grades 4 and 8. Most of the states included in our CITS analyses have six to eight waves of NAEP data prior to the 2010 adoption of CCR standards and four waves of NAEP data after the adoption. The NAEP data are well suited for our CITS analyses because NAEP provides a common set of measures across states and across years. Although NAEP was not designed specifically to be aligned with CCR standards or the CCSS, research that examined the alignment between the NAEP item pool and the CCSS found substantial overlap between the two (Daro et al., 2015). For example, of the items on the 2015 NAEP Grade 4 math assessment, 79% are covered by the CCSS for Grade 4 or below (87% for Grade 8 math). Thus, if the adoption of more rigorous CCR standards had positive effects on student achievement, we would expect the effects to manifest in a larger improvement in NAEP scores even though the NAEP assessments are not perfectly aligned with CCR standards.

Since one goal of this study is to examine whether the effects of states' adoption of more rigorous standards varied by student subgroup, we analyzed state average NAEP scores for all students as well as scores for key student subgroups including students with disabilities (SWDs), English language learners (ELLs), racial/ethnic groups, and students eligible for free or reduced-price lunch (FRPL). NAEP data for subgroups defined by race/ethnicity date back to 1990. NAEP data for students eligible for free- or reduced-priced lunch, however, are not available until 1996, and NAEP data for SWDs and ELLs are not available until 1998.

In addition to NAEP composite scores, we also analyzed NAEP scores for the two reading subscales (i.e., *gaining information* and *literary experience*) and five math subscales (i.e., *algebra; data analysis, statistics, and* *probability; geometry; measurement*; and *number properties and operation*). The rationale for conducting this set of analyses is that the effects of states' adoption of more rigorous CCR standards on student achievement might differ for different NAEP subscales due to uneven alignment between NAEP and CCR standards in different domains within the same subject area. The NAEP alignment study conducted by Daro et al. (2015), for example, revealed that the percentage of items on the 2015 NAEP Grade 4 math assessment covered by the CCSS for Grade 4 or below varies widely across the five math subscales—ranging from 47% to 96%.[8] It is reasonable to assume that the effects of states' adoption of more rigorous standards on NAEP scores would be stronger for NAEP subscales that are more closely assigned with CCR standards than for subscales not as closely aligned.

Measures of Time-Varying Covariates

Our CITS analyses include a set of time-varying covariates to control for potential selection bias and to improve the precision of the treatment effect estimates. A key covariate is the NAEP exclusion rate. Prior to the 1998 NAEP reading administration and 2000 NAEP math administration, NAEP did not allow accommodations for SWDs or ELLs, which resulted in the exclusion of some students who could not meaningfully participate in the assessment without accommodations. To ensure that the NAEP sample be as representative as possible, beginning with the 2002 assessments, NAEP has offered accommodations to all students who need them to demonstrate their knowledge. In the transition years (1998 for reading and 2000 for math), a split sample design was used, with one sample taking the assessment with accommodations and one sample without. Provision of accommodations was found to result in higher levels of inclusion, but with little effect on NAEP scale scores at the national level. Studies of the impact of No Child Left Behind on student achievement conducted by Dee and Jacob (2011) and Wong et al. (2015) also found that their results were not sensitive to whether the analysis was based on NAEP data with accommodations or without accommodations from the transition years. Therefore, we used NAEP data with accommodations from these years in our analyses and also incorporated the NAEP exclusion rate as a time-varying covariate in our CITS model.

Another time-varying covariate included in our CITS analyses is the state-level per-pupil expenditure from the Common Core of Data, which was shown to affect the estimated effects of No Child Left Behind on student achievement (Dee et al., 2013). To account for inflation over time, we used per-pupil expenditure measured in 2016 constant dollars in the CITS analyses. Three additional time-varying covariates included in our CITS analyses are the percentage of students eligible for FRPL, the percentage of non-White students, and pupil-to-teacher ratio, which are also state-level measures from the Common Core of Data. Finally, our CITS analyses include state

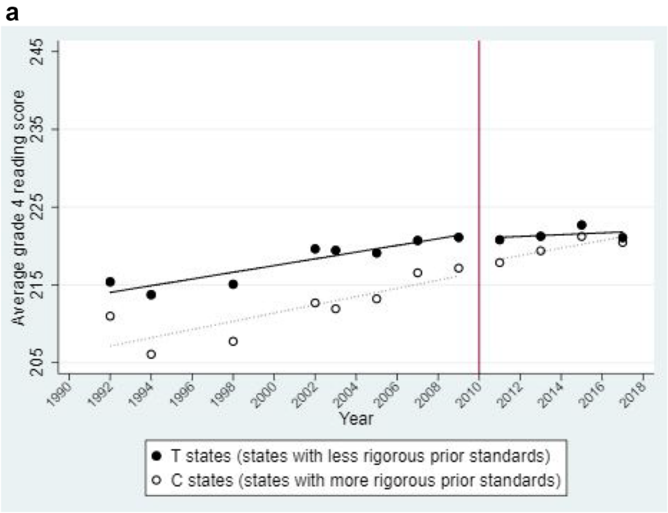

Figure 1. **NAEP achievement trends before and after the adoption of CCR standards in treatment and comparison states defined based on the prior rigor index.**
Figure 1a. **Grade 4 reading.**

Note. N = 17 T states and 12 C states. T = treatment; C = comparison.

unemployment rate as an additional time-varying covariate, the data for which were obtained from the U.S Bureau of Labor Statistics.

Analytic Methods

To inform the modeling choice for our CITS analyses, we examined the pre-CCR trends in both treatment and comparison states, as the validity of a CITS analysis hinges critically on whether the pretreatment trend in each group has a clear functional form and can be easily modeled (Hallberg et al., 2018; St. Clair et al., 2016). Figures 1a–1d present the observed NAEP achievement trends for Grade 4 reading, Grade 8 reading, Grade 4 math, and Grade 8 math, respectively, before and after the adoption of CCR standards in treatment and comparison states defined based on the prior rigor index.[9] It is reassuring that in each case, the pre-CCR trends in both treatment and comparison groups are approximately linear with similar albeit not identical slopes in the two groups, which supports the validity of a CITS linear baseline trend model. Pre-CCR trends for treatment and comparison states defined based on the CCSS-similarity index closely resemble the NAEP math achievement trends shown in Figures 1c and 1d, and their graphic depictions can be found in Supplemental Exhibits 2a and 2b in the online version of the journal.

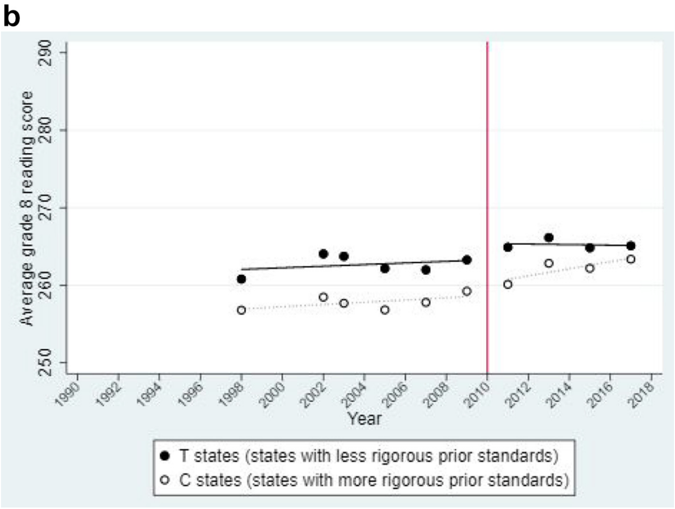

Figure 1b. **Grade 8 reading.**

Note. N = 17 T states and 12 C states. T = treatment; C = comparison.

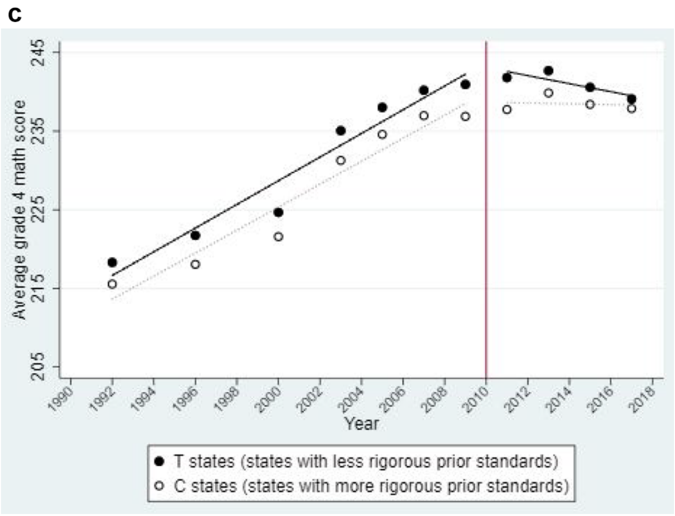

Figure 1c. **Grade 4 math.**

Note. N = 17 T states and 12 C states. T = treatment; C = comparison.

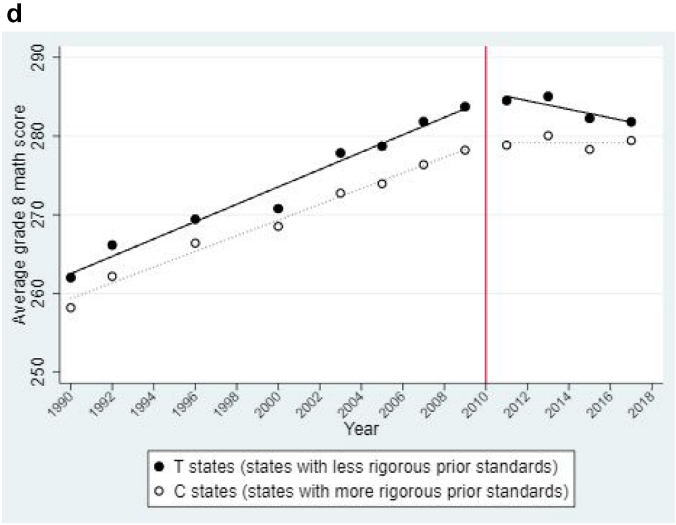

Figure 1d. **Grade 8 math.**

Note. N = 20 T states and 14 C states. T = treatment; C = comparison.

CITS Linear Baseline Trend Model

As specified below, our CITS linear baseline trend model assesses the effects of states' adoption of more rigorous standards on student achievement by comparing the change in student achievement trend from before to after the adoption of CCR standards in the treatment states with the corresponding change in the comparison states in a given analysis sample. The model allows the baseline achievement trend to differ between the treatment and comparison states, and controls for state fixed effects as well as a set of time-varying covariates.

$$Y_{ts}=\sum_{k=1}^{K}\beta_{0k}S_{ks}+\beta_1 TIME_t+\beta_2(T_s*TIME_t)+\sum_{n=1}^{7}\beta_{3n}POST\_YR_{nt}$$
$$+\sum_{n=1}^{7}\beta_{4n}(T_s*POST\_YR_{nt})+\sum_{g=1}^{6}\beta_{5g}X_{gts}+r_{ts}$$

where

- Y_{ts} is the average NAEP score in year t in state s;
- S_{ks}, k = 1, 2, . . ., and K, is a set of dummy indicators for the K states included in the analysis;

- $TIME_t$ is a continuous measure of time measured as the number of years since the first year state NAEP test for a given subject and grade was administered (for Grade 4 reading, $Time_t$ = 0 for year 1992, 2 for 1994, 6 for 1998, . . . and 25 for 2017);
- $T_s * TIME_t$ is an interaction between a state's treatment status and time;
- $POST\_YR_{nt}$, n = 1, 3, 5, and 7, is a set of dummy indicators for the 4 post-CCR NAEP testing years examined ($POST\_YR_{nt}$= 1 for the nth post-CCR year and 0 otherwise);
- $T_s * POST\_YR_{nt}$, n = 1, 3, 5, and 7, is a set of interactions between treatment status and the dummy indicator for each of the 4 post-CCR NAEP testing years;
- X_{gts}, g = 1 ~ 6, is a vector of six time-varying covariates for year t and state s; and
- r_{ts} is a random error associated with year t and state s.

The estimate of primary interest from the above model is β_{4n}, which captures the treatment effect on state average NAEP score in each of the 4 post-CCR NAEP testing years included in the analysis (i.e., 1-, 3-, 5-, and 7-year effects). We estimated the model separately by subject (reading and math) and grade (4 and 8), for NAEP composite scores and subscale scores, and for all students and key student subgroups. For all analyses, the standard errors of the treatment effects were estimated using the block bootstrap method to account for the serial autocorrelation in the time series data.

Robustness Checks

The CITS analyses described above are quasi-experimental in nature and thus subject to threats to internal validity. The validity of the treatment effect estimates from these analyses relies on the assumption that the post-CCR deviation from the pre-CCR achievement trend in the comparison states provides a valid counterfactual for what would have happened in the treatment states had the states not adopted CCR standards. This assumption, however, might not hold if unobserved forces (e.g., changing demographics or economic conditions) occurred during the post-CCR time period and affected student achievement differently in treatment and comparison states. If, for instance, the economic conditions experienced a larger improvement during the post-CCR period in the treatment states relative to the comparison states, then the differential change in the economic conditions of the two groups of states may pose a "history" threat to the internal validity of our CITS analyses, which, if unaccounted for, may potentially result in an overestimated treatment effect.

While it is not possible to rule out all possible threats to internal validity, we checked the robustness of our CITS estimates to some potential internal validity threats following the method used by Dee and Jacob (2011). Specifically, we estimated a model that is similar to our main CITS model but uses a time-varying measure of a state characteristic (e.g., per-pupil expenditure or percentage of students eligible for FRPL) as the dependent variable. A lack of treatment effect on such a measure would rule it out as a potential confounder of the treatment effect on student achievement based on our CITS analyses. A significant treatment effect on such a measure, on the other hand, would speak to the need for incorporating it as a time-varying covariate in the CITS analysis to control for its influence on the treatment effect estimates.

While one would not expect time-varying covariates such as NAEP exclusion rate and measures of school demographic composition to be affected by the treatment in our study (i.e., adoption of more rigorous standards), it is conceivable that covariates such as per-pupil expenditure and pupil-to-teacher ratio may have been affected by the treatment. Given the concern that some of the covariates may potentially be endogenous, we also estimated the CITS model without the covariates as sensitivity analyses.

As another type of robustness checks, we examined the sensitivity of our results to alternative measures of treatment status. In our main CITS analyses, we excluded states with a score of 4 on the 0 to 7 prior rigor index and states in the middle category of the 5-category prior CCSS-similarity index to create a sharper treatment contrast. In one set of sensitivity analyses, we added those previously excluded states to either the treatment group or the comparison group of each analysis sample depending on which option would result in a more balanced sample allocation, and then reestimated the treatment effects. In a second set of sensitivity analyses, we added the previously excluded states to each analysis sample and reestimated the CITS model using the 0–7 prior rigor index and 1–5 prior CCSS-similarity index as continuous measures of treatment strength. The assumption is that the lower the rigor of a state's prior standards as indicated by the prior rigor index, and the more different a state's prior standards were from the CCSS as indicated by the prior CCSS-similarity index, the larger the increase in the rigor of state standards induced by the adoption of the new CCR standards, and hence the larger the expected treatment effects. Finally, as CITS models assuming linear baseline trends can be sensitive to the number of baseline time points included, we also conducted a set of sensitivity analyses that excluded years prior to 2000 from the pre-CCR time period included in our main CITS analyses.

Findings

In this section, we report findings for the effects of states' adoption of more rigorous standards as part of standards-based reform. We first present findings based on CITS analyses of NAEP composite scores and subscale scores in reading and mathematics for all students, and then present findings for select key student subgroups in the NAEP sample. We also briefly summarize findings from robustness checks.

Effects on All Students Based on NAEP Composite Scores

Table 1 presents the estimated effects of states' adoption of more rigorous standards based on CITS analyses in which the treatment and comparison states were defined based on the prior rigor index and the prior CCSS-similarity index of the quality of each state's prior standards, respectively. The table presents separate estimates for effects 1 year, 3 years, 5 years, and 7 years after the adoption of CCR standards (i.e., 1-/3-/5-/7-year effect), in both the original 0–500 NAEP scale and the *SD* unit (i.e., as effect sizes). Contrary to our expectation, the results reveal significant negative effects for Grade 4 reading, with effect sizes ranging from −0.10 to −0.06 across years, which were either significant at the .05 level or marginally significant at the .10 level. Negative effects were also observed for Grade 8 reading, Grade 4 math, and Grade 8 math, although none of those effects were statistically significant except for the 7-year effect for Grade 8 math with state classification based on the prior rigor index (effect size = −0.10, $p < .05$).[10]

To graphically illustrate the CITS analysis results, we plotted the average observed (unadjusted) NAEP scores for the treatment states included in each grade- and subject-specific analysis both before and after the 2010 adoption of CCR standards (shown by the solid black lines in Figures 2a–2f), and their predicted scores 1 year, 3 years, 5 years, and 7 years after adoption had they not adopted more rigorous standards (shown by the dotted lines). The predicted score for the treatment states 1 year after adoption (i.e., 2011) was calculated by subtracting the estimated 1-year effect from the observed 2011 score for the treatment states; the predicted scores for the treatment states in later years were calculated in the same way. The dotted lines in Figures 2a–2f thus show what the NAEP scores would have been for the treatment states in the post-CCR years had they not adopted the new standards—that is, the predicted counterfactual. In the figures, we also display the 95% confidence interval for the treatment effect for each post-CCR year. If the confidence interval does not cover the observed score for the year, it would mean that the corresponding treatment effect is statistically significant at the .05 level. As is clear from Figure 2a, the Grade 4 reading achievement in the treatment states would have improved significantly more after the adoption of the new standards had the states not adopted more rigorous standards, thus reflecting a negative treatment effect. The treatment effects on Grade 8 reading achievement and on Grade 4 and Grade 8 math achievement also tended to be in the negative direction, but of a smaller magnitude, as illustrated in Figures 2b to 2f.

Effects on All Students Based on NAEP Subscale Scores

In addition to NAEP composite scores, we also analyzed NAEP subscales to explore whether the effects of states' adoption of more rigorous standards might differ for different subscales due perhaps to uneven alignment between

Table 1

Estimated Effects of States' Adoption of More Rigorous Standards on Student Achievement as Measured by NAEP Composite Scores, by Subject, Grade, and Timing of Effect

Grade/Subject	Timing of Effect (year after adoption)	No. of States	No. of Observations	Estimate	Standard Error	Effect in *SD*
State classification based on the prior rigor index						
Grade 4 reading	1 year	29	336	−2.22	0.89	−0.06*
Grade 4 reading	3 years	29	336	−3.09	1.28	−0.08*
Grade 4 reading	5 years	29	336	−3.56	1.93	$-0.10^{\dagger}$
Grade 4 reading	7 years	29	336	−3.71	1.91	$-0.10^{\dagger}$
Grade 8 reading	1 year	29	281	0.47	0.85	0.01
Grade 8 reading	3 years	29	281	−1.14	1.08	−0.03
Grade 8 reading	5 years	29	281	−1.48	1.64	−0.04
Grade 8 reading	7 years	29	281	−1.88	1.72	−0.05
Grade 4 math	1 year	34	357	0.29	1.04	0.01
Grade 4 math	3 years	34	357	−0.93	1.28	−0.03
Grade 4 math	5 years	34	357	−1.68	1.72	−0.06
Grade 4 math	7 years	34	357	−2.37	1.85	−0.08
Grade 8 math	1 year	34	381	0.20	1.04	0.01
Grade 8 math	3 years	34	381	−1.12	1.49	−0.03
Grade 8 math	5 years	34	381	−2.43	1.88	−0.07
Grade 8 math	7 years	34	381	−3.88	1.97	−0.10*
State classification based on the prior CCSS-similarity index						
Grade 4 math	1 year	26	271	1.53	1.01	0.05
Grade 4 math	3 years	26	271	0.27	1.42	0.01
Grade 4 math	5 years	26	271	−0.58	1.59	−0.02
Grade 4 math	7 years	26	271	−1.39	1.49	−0.04
Grade 8 math	1 year	26	287	0.85	0.97	0.02
Grade 8 math	3 years	26	287	0.16	1.25	0.00
Grade 8 math	5 years	26	287	−0.42	1.95	−0.01
Grade 8 math	7 years	26	287	−2.46	2.03	−0.06

Note. Effect in *SD* was computed by dividing the estimate in the original NAEP scale by the *SD* of the NAEP test scores for public school students for the relevant grade, subject, and year. *SD* = standard deviation; CCSS = Common Core State Standards.
$^{\dagger}p < .10$. $*p < .05$.

the NAEP assessments and CCR standards in different domains. Table 2 presents the results for the two NAEP reading subscales with state classifications based on the prior rigor index. It shows that the effect estimates for the two reading subscales for Grade 4 were similar in magnitude, and all estimates were either statistically significant ($p < .05$, or $p < .01$) or marginally significant ($p < .10$). For Grade 8, however, there appear to be more notable

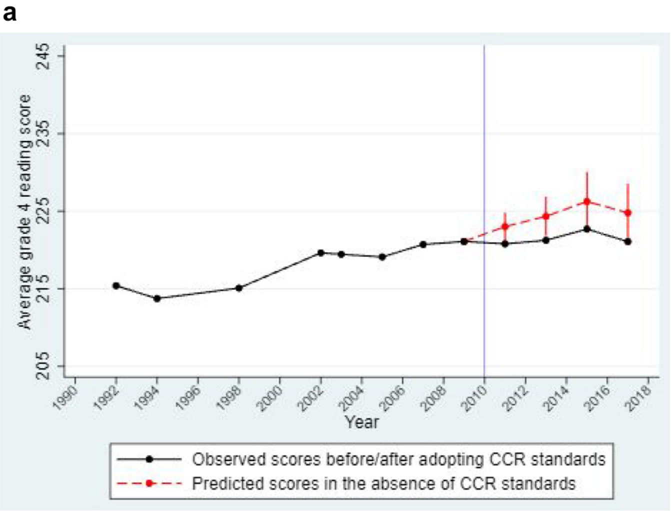

Figure 2. **Observed average NAEP scores for treatment states and their predicted scores in the absence of CCR standards.**

Figure 2a. **Grade 4 reading with state classification based on the prior rigor index.**

Note. Results presented in the figure are based on CITS analysis of data from 17 treatment states and 12 comparison states identified based on the prior rigor index. CCR = college- and career-ready; CITS = comparative interrupted time series.

differences in the results for the two reading subscales. Specifically, while the effect estimates for eighth graders' reading achievement as measured by the *gaining information* subscale during the 7 years after the adoption of CCR standards were all small and nonsignificant (effects = −0.04 to 0.02 *SD*s, $p > .10$), the effect estimates for eighth graders' reading achievement as measured by the *literary experience* subscale were larger, particularly during the period of 3 to 7 years after the adoption of the new standards (effects = −0.09 to −0.08 *SD*s), with the 3-year effect reaching statistical significance ($p < .05$) and the 7-year effect reaching marginal significance ($p < .10$).

Table 3 presents CITS analysis results for NAEP math subscales with state classifications based on the prior rigor index. It shows that for Grade 4, the effects of adopting more rigorous standards on the five NAEP math subscales in each post-CCR year examined were similar in size and all nonsignificant ($p > .10$). For Grade 8, the results for the NAEP math subscales were also similar 1 year and 3 years after the adoption of the new standards, but varied more in later years. While the effect estimates for Grade 8 math differed by 0.05 *SD* or less and were nonsignificant across the five NAEP math subscales in earlier

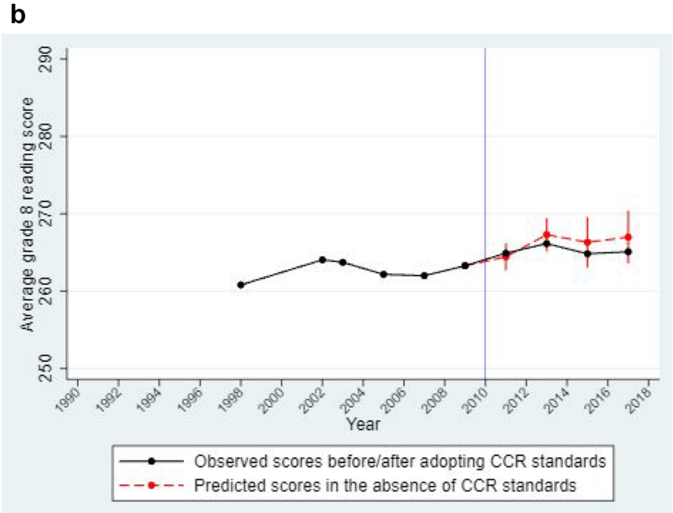

Figure 2b. **Grade 8 reading with state classification based on the prior rigor index.**
Note. Results presented in the figure are based on CITS analysis of data from 17 treatment states and 12 comparison states identified based on the prior rigor index. CCR = college- and career-ready; CITS = comparative interrupted time series.

years, the estimates varied more widely from −0.13 to −0.02 *SD*s 5 years after the adoption of the new standards and from −0.16 to −0.05 *SD*s 7 years after adoption. Relatedly, there were also differences in the statistical significance of the 5- and 7-year effects across the NAEP subscales for Grade 8 math. The effects of adopting more rigorous standards on the *measurement* and *number properties* subscales, for example, were both marginally significant ($p < .10$) 5 years after adoption and statistically significant ($p < .05$) 7 years after adoption. In contrast, the effect estimate for the *algebra* subscale was not significant in any of the years examined ($p > .10$).

Table 4 presents parallel results with state classifications based on the prior CCSS-similarity index. Similar to the results presented in Table 3, most of the effect estimates for the five NAEP math subscales shown in Table 4 were small and nonsignificant. There were some nontrivial differences in the results across different math subscales in certain post-CCR years, but generally the results did not exhibit clear patterns. It is to note that the results presented in Table 4 were based on a substantially smaller sample than the results presented in Table 3, and thus may contain more noise than the results in Table 3.

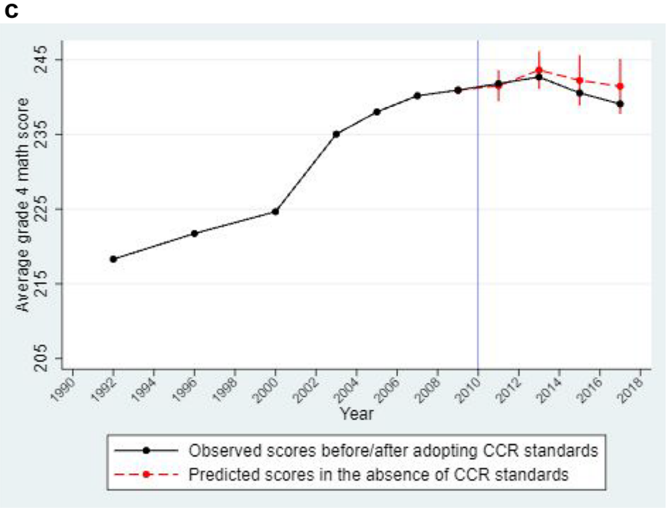

Figure 2c. **Grade 4 math with state classification based on the prior rigor index.**

Note. Results presented in the figure are based on CITS analysis of data from 20 treatment states and 14 comparison states identified based on the prior rigor index. CCR = college- and career-ready; CITS = comparative interrupted time series.

Effects for Student Subgroups

Table 5 presents the effects of states' adoption of more rigorous standards on the achievement of a few key student subgroups, including SWDs, ELLs, Blacks, Hispanics, and students eligible for FRPL. To put these subgroup results in context, we also include in Table 5 the results for all students in the last column. One finding that is clear from the table is that the effect estimates for student subgroups varied more widely than those for the overall sample. While the effect estimates for the overall sample ranged from −0.10 to 0.05 *SD*s across subjects, grades, and years, the estimates varied more widely for student subgroups, particularly for SWDs, ELLs, and Hispanics. The effect estimates for ELLs, for example, ranged from −0.38 to 0.13 *SD*s, with 4 of the 24 estimates having an absolute value exceeding 0.20 *SD*s and 3 of the estimates reaching marginal significance ($p < .10$). The range of effect estimates for SWDs was narrower, but still substantial—from −0.23 to 0.14 *SD*s. In contrast, the effect estimates for students eligible for FRPL had a much narrower range (−0.06 to 0.08 *SD*s), with only two estimates reaching marginal significance.

In addition to effects on student subgroups defined by demographic characteristics, we also examined effects on students performing at different levels

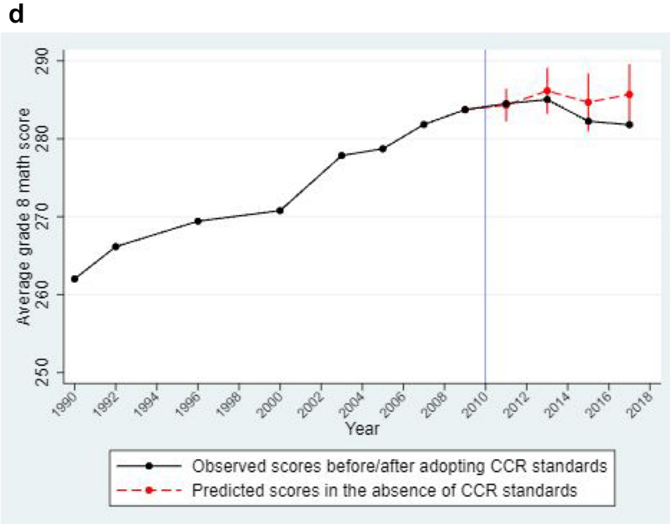

Figure 2d. **Grade 8 math with state classification based on the prior rigor index.**

Note. Results presented in the figure are based on CITS analysis of data from 20 treatment states and 14 comparison states identified based on the prior rigor index. CCR = college- and career-ready; CITS = comparative interrupted time series.

(i.e., 10th, 25th, 50th, 75th, and 90th percentiles) on the NAEP. We found similar patterns of results across students at different levels of performance—most of the effect estimates were negative with a few reaching statistical significance for students at both the lower and upper ends of the NAEP score distribution (see detailed results in the Supplemental Exhibit 3 in the oline version of the journal).

Results From Robustness Checks

To supplement our main CITS analyses, we conducted an extensive set of robustness checks. First, we estimated the "treatment effect" on each time-varying covariate using a model that is similar to our main CITS model but has the covariate as the outcome. Results show that the adoption of more rigorous standards had no statistically significant effect ($p > .05$) in any of the postadoption years examined based on any of the six analytic samples for four of the six time-varying covariates analyzed: per-pupil expenditure, percentage of students eligible for FPRL, percentage of non-White students, and pupil-to-teacher ratio. For NAEP exclusion rate, only one estimate was statistically significant out of the 24 estimates in total associated with each covariate (one estimate for each the four post-CCR years and six analytic

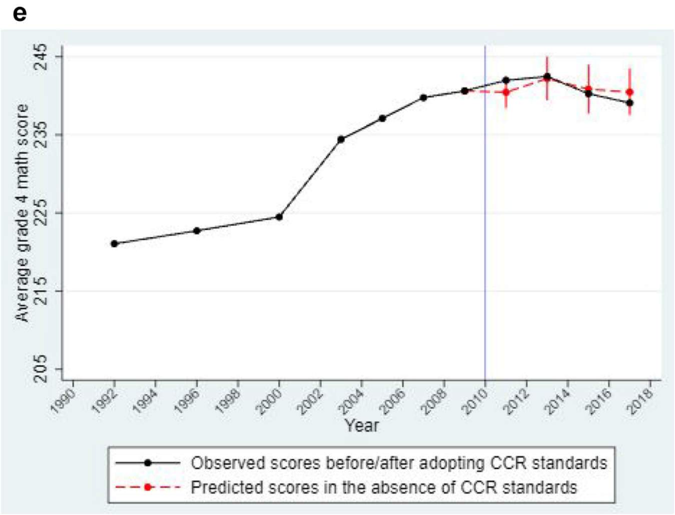

Figure 2e. **Grade 4 math with state classification based on the prior CCSS-similarity index.**

Note. Results presented in the figure are based on CITS analysis of data from 14 treatment states and 12 comparison states identified based on the prior CCSS-similarity index. CCR = college- and career-ready; CITS = comparative interrupted time series; CCSS = Common Core State Standards.

samples). Given the large number of analyses conducted, it is possible that this significant estimate may be due to chance. For state unemployment rate, almost one third (7 out of 24) of the estimates were statistically significant. Our sensitivity analyses, however, show that our CITS results are not sensitive to the inclusion of these covariates—estimates based on CITS models with and without the covariates are very similar with differences less than 0.01 *SD*s in most cases and less than 0.02 *SD*s in all cases.

As another type of robustness checks, we examined the sensitivity of our results to alternative measures of treatment status. We first added to the CITS analysis samples the states excluded from the main CITS analyses (i.e., states in the middle category of the prior rigor index or the prior CCSS-similarity index), and then reestimated the CITS model based on both a dichotomous version and a continuous version of the treatment indicator with each analysis sample. Results from these analyses are largely consistent with the results from the main CITS analyses with minor differences.

In our final set of robustness checks, we restricted the pre-CCR period included in our CITS analyses to more recent years. Once we excluded years

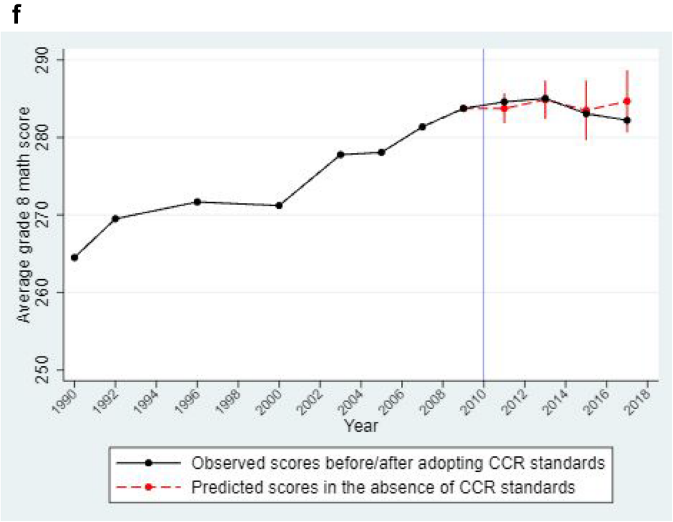

Figure 2f. **Grade 8 math with state classification based on the prior CCSS-similarity index.**

Note. Results presented in the figure are based on CITS analysis of data from 14 treatment states and 12 comparison states identified based on the prior CCSS-similarity index. CCR = college- and career-ready; CITS = comparative interrupted time series; CCSS = Common Core State Standards.

prior to 2000, the effect estimates became less negative and none of the estimates had *p* values less than .10 for reading. For math, however, the effect estimates based on the substantially shortened pre-CCR period became more negative and a few estimates for 7-year effects were statistically significant ($p < .05$) or marginally significant ($p < .10$; see detailed results in Supplemental Exhibit 4 in the oline version of the journal). Nevertheless, the general pattern of results from this set of sensitivity analyses remains the same as that from our main CITS analyses—most of the effect estimates were small, negative, and nonsignificant, but a few negative estimates did reach statistical significance or marginal significance.

Discussion

Relying on state-level NAEP data from the past three decades, this study tested the basic premise of the current wave of standards-based reform—that is, the adoption of more rigorous standards would lead to improved student achievement. Contrary to our expectation, we found that states' adoption of

Table 2

Estimated Effects of States' Adoption of More Rigorous Standards on Student Achievement as Measured by NAEP Reading Subscales, by Grade and Timing of Effect (With State Classification Based on the Prior Rigor Index)

		Grade 4			Grade 8		
Reading Subscale	Timing of Effect	Estimate	*SE*	Effect in *SD*	Estimate	*SE*	Effect in *SD*
Gaining information	1 year	−2.11	1.00	−0.06*	0.89	0.88	0.02
Literary experience	1 year	−2.37	0.88	−0.07**	−0.07	0.93	−0.01
Gaining information	3 years	−2.70	1.29	−0.07*	−0.18	1.08	−0.01
Literary experience	3 years	−3.53	1.33	−0.10**	−2.44	1.22	−0.08*
Gaining information	5 years	−3.07	1.81	$-0.08^{\dagger}$	−0.63	1.69	−0.02
Literary experience	5 years	−4.07	2.16	$-0.11^{\dagger}$	−2.79	1.70	−0.08
Gaining information	7 years	−3.53	1.99	$-0.09^{\dagger}$	−1.29	1.77	−0.04
Literary experience	7 years	−3.94	1.96	−0.11*	−3.13	1.79	$-0.09^{\dagger}$

Note. N = 29 states and 339 observations for Grade 4 analyses; *N* = 29 states and 281 observations for Grade 8 analyses. Effect in *SD* was computed by dividing the estimate in the original NAEP scale by the *SD* of the NAEP test scores for public school students for the relevant grade, subject, and year. *SE* = standard error; *SD* = standard deviation.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

more rigorous standards had significant negative effects on students' reading achievement during the first 7 years after adoption for fourth graders, but not for eighth graders. In math, we found that states' adoption of more rigorous standards did not have any significant effect during the time period examined for fourth graders, but had a significant negative effect for eighth graders 7 years after the adoption of the more rigorous standards. The size of these significant effects, however, was quite modest, ranging from −0.10 to −0.06 *SD*s. Moreover, the negative 7-year effect on eighth graders' math achievement was significant only when state classification was based on the prior rigor index; it was smaller and not significant when state classification was based on the prior CCSS-similarity index.

Results from sensitivity analyses based on alternative measures of treatment status are largely consistent with the above results from our main CITS analyses. Results from sensitivity analyses based on a much shorter pre-CCR period, however, tended to be less negative for reading but more negative for math, compared with the main CITS results. Nevertheless, the general pattern of results remains the same across the various types of sensitivity analyses—most of the effect estimates were small, negative, and nonsignificant, but a few negative estimates did reach statistical significance or marginal significance.

Table 3

Estimated Effects of States' Adoption of More Rigorous Standards on Student Achievement As Measured by NAEP Math Subscales, by Grade and Timing of Effect (With State Classification Based on the Prior Rigor Index)

		Grade 4			Grade 8		
Math Subscale	Timing of Effect	Estimate	*SE*	Effect in *SD*	Estimate	*SE*	Effect in *SD*
Algebra	1 year	0.42	0.92	0.01	0.55	1.02	0.02
Data analysis	1 year	0.85	1.28	0.03	0.96	1.19	0.03
Geometry	1 year	0.64	1.11	0.02	0.19	1.10	0.01
Measurement	1 year	0.99	1.16	0.03	−0.44	1.50	−0.01
Number properties	1 year	−0.41	1.16	−0.01	−0.49	0.97	−0.01
Algebra	3 years	−0.69	1.03	−0.02	−0.52	1.44	−0.01
Data analysis	3 years	0.85	1.64	0.03	−0.35	1.84	−0.01
Geometry	3 years	0.00	1.54	0.00	−1.18	1.67	−0.03
Measurement	3 years	−1.28	1.69	−0.04	−2.15	2.04	−0.06
Number properties	3 years	−1.65	1.44	−0.06	−1.77	1.34	−0.05
Algebra	5 years	−1.15	1.40	−0.04	−0.79	1.85	−0.02
Data analysis	5 years	−2.01	2.00	−0.07	−2.44	2.15	−0.07
Geometry	5 years	−1.74	2.11	−0.06	−2.64	2.18	−0.07
Measurement	5 years	−2.05	2.21	−0.07	−4.79	2.58	−0.13†
Number properties	5 years	−1.61	1.89	−0.05	−2.90	1.56	−0.08†
Algebra	7 years	−2.04	1.73	−0.07	−1.99	1.78	−0.05
Data analysis	7 years	−3.41	2.19	−0.11	−4.60	2.46	−0.12†
Geometry	7 years	−3.32	2.30	−0.11	−4.58	2.40	−0.12†
Measurement	7 years	−3.23	2.37	−0.10	−6.43	2.91	−0.16*
Number properties	7 years	−1.46	1.89	−0.05	−3.60	1.56	−0.09*

Note. *N* = 34 states and 357 observations for Grade 4 analyses; *N* = 34 states and 381 observations for Grade 8 analyses. Effect in *SD* was computed by dividing the estimate in the original NAEP scale by the *SD* of the NAEP test scores for public school students for the relevant grade, subject, and year. *SE* = standard error; *SD* = standard deviation.
$^{\dagger}p < .10$. $^{*}p < .05$.

Analyses of NAEP subscale scores show that the effects of adopting more rigorous standards on the two NAEP reading subscales were similar for Grade 4, but differed for Grade 8, with significant negative effects on eighth graders' performance on the *literary experience* subscale and smaller nonsignificant effects on the *gaining information* subscale. Similarly, we found that the effects of adopting more rigorous standards on the five NAEP math subscales were similar for Grade 4, but differed for Grade 8, particularly in later years (i.e., 5 and 7 years after adoption), according to analyses with state classifications based on the prior rigor index. Analyses with state classifications based on the prior CCSS-similarity index also revealed some nontrivial differences in

Table 4

Estimated Effects of States' Adoption of More Rigorous Standards on Student Achievement as Measured by NAEP Math Subscales, by Grade and Timing of Effect (With State Classification Based on the Prior CCSS-Similarity Index)

		Grade 4			Grade 8		
Math Subscale	Timing of Effect	Estimate	*SE*	Effect in *SD*	Estimate	*SE*	Effect in *SD*
Algebra	1 year	1.95	0.94	0.07*	0.60	1.03	0.02
Data analysis	1 year	2.44	1.41	0.08†	1.65	1.33	0.05
Geometry	1 year	0.99	1.19	0.03	0.04	1.13	0.00
Measurement	1 year	1.87	0.97	0.06†	1.10	1.32	0.03
Number properties	1 year	1.16	1.18	0.04	1.47	0.92	0.04
Algebra	3 years	1.64	1.22	0.05	0.25	1.38	0.01
Data analysis	3 years	2.79	1.92	0.09	0.89	1.80	0.02
Geometry	3 years	−0.24	1.63	−0.01	−0.41	1.50	−0.01
Measurement	3 years	−0.13	1.55	0.00	−0.32	1.61	−0.01
Number properties	3 years	−0.52	1.71	−0.02	0.80	1.29	0.02
Algebra	5 years	1.83	1.38	0.06	1.13	2.02	0.03
Data analysis	5 years	−0.15	2.16	0.00	−1.20	2.52	−0.03
Geometry	5 years	−2.56	2.09	−0.09	−1.17	2.11	−0.03
Measurement	5 years	−1.63	1.92	−0.05	−1.46	2.57	−0.04
Number properties	5 years	−0.38	1.87	−0.01	−0.30	1.72	−0.01
Algebra	7 years	−0.10	1.59	0.00	−0.17	1.96	0.00
Data analysis	7 years	−1.94	2.36	−0.06	−4.07	2.86	−0.10
Geometry	7 years	−3.65	2.07	−0.12†	−4.02	2.10	−0.10†
Measurement	7 years	−2.75	1.73	−0.09	−3.77	2.81	−0.10
Number properties	7 years	−0.26	1.69	−0.01	−1.75	1.78	−0.04

Note. *N* = 26 states and 271 observations for Grade 4 analyses; *N* = 26 states and 287 observations for Grade 8 analyses. Effect in *SD* was computed by dividing the estimate in the original NAEP scale by the *SD* of the NAEP test scores for public school students for the relevant grade, subject, and year. *SE* = standard error; *SD* = standard deviation.
$^{\dagger}p < .10$. $^{*}p < .05$.

effect estimates across different math subscales, but generally the results did not exhibit clear patterns. Finally, our subgroup analyses show that the effects of states' adoption of more rigorous standards for certain student subgroups—SWDs, ELLs, and Hispanics in particular—varied more widely across subjects, grades, and years, and tended to have a much larger negative effect size than the effects for the overall sample, which may be partially due to the smaller sample sizes for these subgroups and thus less stable estimates.

Overall, findings from this study indicate that the adoption of more rigorous standards that has been the foundation of the latest standards-based reform had not yet moved the needle on student achievement during the first

Table 5

Estimated Effects of States' Adoption of More Rigorous Standards on the Achievement of Key Student Subgroups as Measured by NAEP Composite Scores, by Subject, Grade, and Timing of Effect

	Timing of Effect	Effect in *SD*					
Grade/Subject	(year after adoption)	SWDs	ELLs	Blacks	Hispanics	FRPL	All
State classification based on the prior rigor index							
Grade 4 reading	1 year	−0.04	0.00	−0.09*	−0.10†	−0.03	−0.06*
Grade 4 reading	3 years	0.00	0.03	−0.05	−0.05	0.00	−0.08*
Grade 4 reading	5 years	−0.05	0.00	−0.10	−0.16†	−0.03	−0.10†
Grade 4 reading	7 years	−0.04	0.03	−0.11	−0.09	−0.01	−0.10†
Grade 8 reading	1 year	−0.06	0.13	0.07	0.09	0.03	0.01
Grade 8 reading	3 years	−0.13	0.00	0.04	0.00	−0.02	−0.03
Grade 8 reading	5 years	−0.15	−0.13	0.00	0.01	−0.01	−0.04
Grade 8 reading	7 years	−0.23†	−0.13	−0.01	0.01	0.00	−0.05
Grade 4 math	1 year	0.13*	0.06	0.06	0.05	0.07†	0.01
Grade 4 math	3 years	0.05	−0.03	0.02	0.03	0.02	−0.03
Grade 4 math	5 years	0.03	−0.07	0.01	0.03	0.00	−0.06
Grade 4 math	7 years	0.05	−0.19	−0.05	0.02	−0.01	−0.08
Grade 8 math	1 year	0.02	−0.18	0.02	−0.08	0.03	0.01
Grade 8 math	3 years	−0.09	−0.11	−0.04	−0.03	0.00	−0.03
Grade 8 math	5 years	−0.12	−0.34	−0.08	−0.08	−0.03	−0.07
Grade 8 math	7 years	−0.18	−0.38†	−0.11	−0.11	−0.06	−0.10*
State classification based on the prior CCSS-similarity index							
Grade 4 math	1 year	0.14*	−0.03	0.06	0.05	0.08†	0.05
Grade 4 math	3 years	0.04	−0.20†	0.02	0.02	0.03	0.01
Grade 4 math	5 years	−0.01	−0.26	−0.03	−0.03	−0.02	−0.02
Grade 4 math	7 years	−0.01	−0.32†	−0.08	−0.06	−0.05	−0.04
Grade 8 math	1 year	−0.02	−0.03	0.01	−0.11†	0.04	0.02
Grade 8 math	3 years	−0.08	0.06	0.01	−0.08	0.04	0.00
Grade 8 math	5 years	−0.12	0.03	−0.06	−0.11	0.02	−0.01
Grade 8 math	7 years	−0.23†	−0.04	−0.10†	−0.18*	−0.02	−0.06

Note. See Supplemental Exhibit 5 (in the online version of the journal) for the sample size for each subgroup analysis. *SD* = standard deviation; SWDs = students with disabilities; ELLs = English language learners; FRPL = free- or reduced-price lunch; CCSS = Common Core State Standards.

$^{\dagger}p < .10$. $^{*}p < .05$.

7 years after adoption—a conclusion resonating with the conclusion from earlier analyses conducted by Loveless (2014, 2015, 2016) about the impact of the CCSS. What is not clear from this study though is why. Below we reflect on some of the issues that may affect the interpretation of the findings from this study, including study limitations. We then discuss implications of the

findings for researchers studying standards-based reforms and for education policy makers and administrators who play an important role in supporting such reforms.

Study Limitations and Other Potential Explanations for Study Findings

This study has a number of limitations. First, given the timing of CCR standards adoption across states, a true "no-treatment" comparison group is not available for this study. Instead, we constructed the treatment and comparison groups for an CITS design based on the natural variation in the quality of states' prior standards among states that adopted CCR standards in 2010. The CITS analyses thus would produce unbiased estimates of the effects of adopting more rigorous standards for the treatment states only under certain conditions—that is, if the adoption of CCR standards had no effect on student achievement in the comparison states whose prior standards were already fairly rigorous, and if factors unaccounted for in our CITS model affected student achievement in treatment and comparison states in similar ways. If the adoption of more rigorous standards had a positive effect on student achievement in the comparison states, then our CITS estimates would provide a lower bound of the true effect estimates. Conversely, if the adoption of more rigorous standards had a negative effect on student achievement in the comparison states, then our CITS estimates would provide an upper bound of the true effect estimates. Furthermore, by definition, the comparison states had adopted rigorous standards earlier than the treatment states did. We do not know what factors led them to adopt high standards earlier, but it is possible that the same factors might have affected the two groups of states differently in their achievement growth during the post-CCR period, which might have introduced some bias to our CITS estimates.[11]

Related to the selection of treatment and comparison states, another limitation of the study concerns its external validity. The findings from this study pertain to the effects of adopting more rigorous standards on the subset of states defined as treatment states in this study—that is, states whose prior content standards were less rigorous or less like the CCSS. It is possible that the effects of adopting more rigorous standards on the comparison states and states excluded from our CITS analyses may be different.

A third limitation of the study is that our measures of student achievement—NAEP scores—are not perfect measures of students' college and career readiness, the target outcome of CCR standards.[12] The study of the alignment between NAEP and the CCSS, the dominant form of CCR standards, for example, revealed that over 20% of the items on the 2015 NAEP Grade 4 math assessment and 13% of the items on the 2015 NAEP Grade 8 math assessment are not covered by the CCSS for the relevant grade or below (Daro et al., 2015). Therefore, NAEP may be less sensitive to changes in student achievement induced by the new CCR standards than assessments more closely aligned with the new standards. On the other hand, a test perfectly aligned with the new standards may not provide the best test of the effects of adopting the new standards, if the hope is that teaching would not focus narrowly just on what is specified in the standards but aim to improve students' knowledge and skills more broadly.

Fourth, when interpreting the study findings, it is important to bear in mind that the study was designed to estimate the effects of more rigorous standards *as implemented* in the states included in our analyses 1 year, 3 years, 5 years, and 7 years after the adoption of the new standards. Essentially these are analogous to the "intent-to-treat" (ITT) effects that represent the effects of being assigned to the treatment condition in a randomized controlled trial, regardless of whether the treatment was implemented as intended. These ITT effects from our study are not the same as the effects of 1 year, 3 years, 5 years, and 7 years of full implementation of more rigorous standards. As mentioned earlier, most states *expected* to take 3 to 5 years to fully transition from the old to the new standards, and it may take even longer to actually complete the transition. Thus, despite a seemingly fairly long follow-up period (7 years), findings from this study largely reflect early effects of states' implementation of the new CCR standards during the transition period, which may be different from the effects after the new standards were fully implemented with high fidelity. It is also worth noting that the findings from our study represent the effects of adopting—rather than the effects of sustaining—more rigorous state standards because the rigor of the standards actually enacted may have changed over time after the 2010 adoption in both treatment and comparison states.

Given these limitations, findings from our study might either over- or underestimate the true effects of implementing more rigorous standards. But to the extent that our findings reflect true effects, what might account for the lack of positive findings? One explanation lies in the multitude of challenges that states, districts, schools, and teachers experienced during the transition period, as discussed in the literature review section of the article. Given all those challenges, ambitious instructional goals put forth by the new standards may have failed to bring about fundamental changes in teachers' instructional practice—a familiar finding from research on previous standards-based reforms (e.g., Cohen, 1990; Spillane & Zeuli, 1999). Indeed, well-designed standards are essential, but not sufficient, for the success of standards-based reforms. In order for the new standards to ultimately produce meaningful gains in student learning, the new standards must first induce real changes in teacher knowledge and instructional practice, which are key mediators in the pathway from the adoption of the new standards to improved student learning (Cohen & Hill, 2000; Hill & Chin, 2018). As Cohen and Hill (2000) argue, "teachers figure as a key connection between policy and practice, their opportunities to learn about and from policy are a crucial influence both on their practice and, at least indirectly, on student achievement" (p. 294).

Such learning opportunities for teachers are particularly critical for the success of reforms that call for intellectually much more ambitious instruction. Very often, however, standards-based reforms failed to provide teachers with sufficient learning opportunities and supports, and hence led to only modest changes in instructional practice and minimal impact on student achievement (Coburn, 2004; Coburn et al., 2016).

It is harder to arrive at plausible explanations for why the adoption of more rigorous standards may have had a negative effect on student achievement as revealed by some of our analyses. Perhaps more rigorous standards might harm student achievement if, for example, they were overly challenging, ill-suited to students' readiness level, and took students out of their zone of proximal development. It is also conceivable that many teachers were not well prepared to enact intellectually more ambiguous instruction called for by the more rigorous CCR standards and as a result taught the new standards poorly and in a less engaging way, which may have led to less student learning compared with old standards that were less rigorous but taught well and in a more engaging way.

Implications for Future Research and Policy

While speculations about the potential explanations for the results from our study abound, the relative importance of these explanations is unclear. What is clear is that the results of this study do not support the hypothesis that adopting more rigorous standards would lead to significant improvement in student achievement. A useful line of future research is to examine the extent to which states' adoption of more rigorous standards have led to improvement in teacher knowledge (both subject knowledge and pedagogical knowledge) and instructional practice aligned with the standards. A clear understanding about the impact of the new standards on what teachers know and how they teach in the classrooms will help inform our understanding of the impact of the standards on student learning. It also will help identify weak or broken link(s) in the pathway between the adoption of new standards and student outcomes, and inform efforts to strengthen the linkage between different steps along the pathway. Relatedly, it would be useful to know how curricular materials and assessments have changed in response to the new standards. Furthermore, given the unique nature of the "treatment," research on standards-based reforms should take a long-term view with an extended timeline. Building on the study reported in this article, further follow-up research is needed to fully capture the impact of states' adoption of more rigorous standards as states complete the transition process and as the implementation of the new standards widens and deepens over time.

For education policymakers and administrators, the findings from this study serve as a reminder that well-crafted standards do not automatically translate to improved student test scores. Statewide transition of academic standards is a massive undertaking and a highly challenging endeavor that requires concerted efforts and support at multiple levels based on a coherent, systemic approach (Smith & O'Day, 1991). Although well-crafted standards are arguably the centerpiece of standards-based reform, they are just one piece of the puzzle, and the other pieces also need to be in place, supporting and reinforcing one another, in order for the standards to achieve their intended impact on teaching and learning. Of particular importance are curricular, instructional, and assessment materials well aligned to the new standards, as well as rich support for teachers—the frontline implementers of the standards—to learn about the new standards, buy into the standards, and make fundamental shifts in their instructional practice in accordance with the new standards. When and to what extent the potential of the new standards will be realized will depend on how quickly and how adequately all the needed supporting pieces can be put in place.

Appendix

Year of CCR Standards Adoption, Quality of Prior Content Standards, and Treatment Status of Each State

			Prior Rigor Index for Prior ELA Standards		Prior Rigor Index for Prior Math Standards		Prior CCSS-Similarity Index for Prior Math Standards	
State	Year CCR Standards in ELA Adopted	Year CCR Standards in Math Adopted	Score	Treatment	Score	Treatment	Score	Treatment
Alabama	2010	2010	6	C	5	C	5	C
Alaska[a]	2012	2012	1	NA	3	NA	3	NA
Arizona	2010	2010	5	C	4	NA	1	T
Arkansas	2010	2010	3	T	3	T	3	NA
California	2010	2010	7	C	7	C	5	C
Colorado	2009	2009	6	NA	3	NA	3	NA
Connecticut	2010	2010	2	T	3	T	2	T
DC	2010	2010	7	C	7	C	NA	NA
Delaware	2010	2010	2	T	5	C	3	NA
Florida	2010	2010	5	C	7	C	5	C
Georgia	2010	2010	6	C	6	C	5	C
Hawaii	2010	2010	4	NA	3	T	3	NA
Idaho	2011	2011	4	NA	5	NA	4	NA
Illinois	2010	2010	3	T	1	T	2	T
Indiana	2010	2010	7	C	7	C	5	C
Iowa	2010	2010	1	T	3	T	1	T
Kansas	2010	2010	4	NA	1	T	1	T
Kentucky	2010	2010	3	T	2	T	1	T
Louisiana	2010	2010	6	C	3	T	1	T
Maine	2011	2011	4	NA	3	NA	2	NA
Maryland	2010	2010	4	NA	3	T	2	T

(continued)

Appendix **(continued)**

State	Year CCR Standards in ELA Adopted	Year CCR Standards in Math Adopted	Prior Rigor Index for Prior ELA Standards		Prior Rigor Index for Prior Math Standards		Prior CCSS-Similarity Index for Prior Math Standards	
			Score	Treatment	Score	Treatment	Score	Treatment
Massachusetts	2010	2010	7	C	6	C	3	NA
Michigan	2010	2010	2	T	6	C	5	C
Minnesota[a]	2010	2007	4	NA	5	NA	5	NA
Mississippi	2010	2010	3	T	4	NA	5	C
Missouri	2010	2010	3	T	2	T	2	T
Montana	2011	2011	2	NA	0	NA	2	NA
Nebraska[a]	2014	2015	1	NA	3	NA	2	NA
Nevada	2010	2010	4	NA	4	NA	1	T
New Hampshire	2010	2010	4	NA	3	T	2	T
New Jersey	2010	2010	4	NA	4	NA	1	T
New Mexico	2010	2010	4	NA	4	NA	3	NA
New York	2010	2010	3	T	5	C	3	NA
North Carolina	2010	2010	3	T	3	T	3	NA
North Dakota	2011	2011	2	NA	4	NA	4	NA
Ohio	2010	2010	4	NA	3	T	3	NA
Oklahoma	2010	2010	5	C	5	C	5	C
Oregon	2010	2010	4	NA	5	C	4	C
Pennsylvania	2010	2010	3	T	1	T	3	NA
Rhode Island	2010	2010	3	T	3	T	1	T
South Carolina	2010	2010	3	T	3	T	3	NA
South Dakota	2010	2010	4	NA	3	T	4	C
Tennessee	2010	2010	6	C	3	T	4	C
Texas[a]	2008	2008	6	NA	4	NA	3	NA
Utah	2010	2010	4	NA	6	C	4	C
Vermont	2010	2010	2	T	1	T	3	NA
Virginia[a]	2010	2009	6	C	4	NA	2	NA
Washington	2011	2011	4	NA	7	NA	5	NA
West Virginia	2010	2010	3	T	5	C	3	NA
Wisconsin	2010	2010	3	T	1	T	1	T
Wyoming	2012	2012	3	NA	1	NA	2	NA

Note. The prior rigor index for a state's 2010 content standards is on a 0–7 point scale, with 7 presenting the highest rigor (Carmichael et al., 2010). For our main CITS analyses with state classification based on the prior rigor index, treatment states are states with a score of 0–3 and comparison states are states with a score of 5–7 on the index. States with a score of 4 on the index were excluded from the main CITS analyses but included in sensitivity analyses. The original measure of the similarity between a state's 2009 content standards and the CCSS in mathematics is on a 0–1000 point scale (Schmidt & Houang, 2012). For this study, we used a 1–5 version of the measure, with 1 representing "least like CCSS" and 5 "most like CCSS," based on the Schmidt and Houang's categorization. For our main CITS analyses with state classification based on the prior CCSS-similarity index, treatment states are states in the two "least like CCSS" groups, and comparison states are states in the two "most like CCSS" groups. States in the middle group were excluded from the main CITS analyses but included in sensitivity analyses. Columns with the heading "Treatment" show the treatment status of each state in the CITS analysis with state classifications based on the given index: T = treatment state; C = comparison state; NA = excluded from main CITS analysis; CCR = college- and career-ready; ELA = English language arts; CCSS = Common Core State Standards.

[a] Alaska, Nebraska, Texas, and Virginia adopted their own CCR standards in both ELA and mathematics. Minnesota adopted the CCSS in ELA but not math. All other states adopted the CCSS in both subjects.

Notes

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C150007 to University of Pennsylvania—Graduate School of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

[1] The 45 states include Minnesota which adopted the CCSS only in ELA but not in math. Four states (Alaska, Nebraska, Texas, and Virginia) did not adopt the CCSS, but adopted their own CCR standards between 2008 and 2015.

[2] For simplicity, we will count District of Columbia (DC) as one of 51 states hereinafter.

[3] Carmichael et al. (2010) also analyzed the CCSS and gave it a rating of 7 in math and 6 in ELA on the 0 to 7 scale for the content and rigor of the standards. The study concluded that the CCSS were clearly superior to the standards in the vast majority of states.

[4] Schmidt and Houang (2012) reported that the codings of standards documents in TIMSS based on the same procedures achieved reliabilities of .70 or higher. Furthermore, the prior CCSS-similarity index for math and the prior rigor index for math are strongly correlated across the 50 states (correlation = .67, $p < .001$).

[5] Comparisons of Virginia's own version of CCR standards with the CCSS show that the two sets of standards are strongly aligned overall, with the main difference between the two being in the organization and learning progressions rather than in content and rigor (see http://www.doe.virginia.gov/testing/common_core/index.shtml for details).

[6] This information is based primarily on the report by Norton et al. (2016), supplemented with web searches. The four states that had repealed the CCSS by late 2016 include Arizona, Indiana, Oklahoma, and South Carolina. An analysis of Oklahoma's new CCR standards conducted by Achieve (2016) concluded that the new standards adopted by Oklahoma to replace the CCSS were akin to their pre-CCSS standards. The new CCR standards that replaced the CCSS in the other three states were all very similar to the CCSS (Achieve, 2015; Cavazos, 2018; Hinton, 2016).

[7] For states that adopted CCR standards in odd-numbered years (i.e., 2007, 2009, 2011, and 2015), the available NAEP data could only be used to estimate 2-, 4-, 6-, and/or 8-year effects. The number of such states is too small to allow for reliable estimation. (See the appendix for the timing of CCR standards adoption in each state.)

[8] For Grade 4, the specific percentages for the five NAEP math subscales are as follows: 62% for *algebra;* 47% for *data analysis, statistics, and probability*; 68% for *geometry*; 96% for *measurement*; and 90% for *number properties and operation.* For Grade 8, the corresponding percentages are 84%, 74%, 80%, 100%, and 97%, respectively (Daro et al., 2015). Similar information is not available for NAEP reading subscales.

[9] One may note from Figure 1 that states with less rigorous prior standards (treatment states) had higher pre-CCR NAEP scores than states with more rigorous prior standards (comparison states). One likely explanation is that treatment states in our study tended to have somewhat more advantaged student populations than comparison states. The average percentage of minority students during the pre-CCR years, for example, was 30.3% for treatment states versus 38.9% for comparison states for the CITS reading analyses.

[10] The statistical significance for the four effects for Grade 4 reading remained unchanged even after we applied corrections for multiple comparisons within grade and subject using the Benjamini-Hochberg method used by the What Works Clearinghouse (2020). The significant 7-year effect for Grade 8 math, however, was no longer significant after the correction.

[11] Since comparison states in our study already had relatively more rigorous pre-CCR standards, it is possible, for example, they might have been better prepared for a smooth transition to the new CCR standards than treatment states.

[12] As an extension of the analyses presented in this paper, we are currently examining the effects of states' adoption of more rigorous standards on a different type of CCR-related outcome–high school graduation rate–using a similar CITS design. We had also planned to examine college enrollment as an additional outcome; however, data on college enrollment do not meet the linearity assumption underlying a valid CITS linear baseline trend model.

References

Achieve. (2015). *A review of the South Carolina college- and career-ready standards in English language arts and mathematics*. https://www.achieve.org/publications/review-south-carolina-college-and-career-ready-standards

Achieve. (2016). *A review of the Oklahoma January 2016 English language arts and mathematics academic standards*. https://www.achieve.org/publications/review-oklahoma-january-2016-english-language-arts-and-mathematics-academic-standards

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping state proficiency standards onto NAEP scales: 2005-2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Bay-Williams, J., Duffett, A., & Griffith, D. (2016). *Common core math in the K–8 classroom: Results from a national teacher survey*. Thomas B. Fordham Institute.

Carmichael, S. B., Martino, G., Porter-Magee, K., & Wilson, W. S. (2010). *The state of the state standards—and the Common Core—in 2010*. Thomas B. Fordham Institute.

Cavazos, S. (2018, August 22). New report from Common Core supporters says Indiana's Academic Standards are almost as good. *Chalkbeat Indiana*. https://in.chalkbeat.org/2018/8/22/21105559/new-report-from-common-core-supporters-says-indiana-s-academic-standards-are-almost-as-good

Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, *77*(3), 211–244. https://doi.org/10.1177/003804070407700302

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, *45*(4), 243–251. https://doi.org/10.3102/0013189X16651080

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*(3), 311–329. https://doi.org/10.3102/01623737012003311

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record*, *102*(2), 294–343. https://doi.org/10.1111/0161-4681.00057

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at Grades 4 and 8 to the Common Core State Standards (CCSS) for mathematics*. American Institutes for Research.

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. https://doi.org/10.1002/pam.20586

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252–279. https://doi.org/10.3102/0162373712467080

Education Week (2017, September 18). *Map: Tracking the Common Core State Standards*. Retrieved from https://www.edweek.org/ew/section/multimedia/map-states-academic-standards-common-core-or.html

Gewertz, C. (2015, September 28). The Common Core explained. *Education Week*. Retrieved from http://www.edweek.org/ew/issues/common-core-state-standards/

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher*, *47*(5), 295–306. https://doi.org/10.3102/0013189X18769302

Heitin, L. (2015, March 17). Review of math programs comes under fire. *Education Week*. https://www.edweek.org/teaching-learning/review-of-math-programs-comes-under-fire/2015/03

Heitin, L. (2016, August 30). Common-core reading materials get mixed results in first major review. *Education Week*. https://www.edweek.org/teaching-learning/common-core-reading-materials-get-mixed-results-in-first-major-review/2016/08

Henderson, M. B., Houston, D. M., Peterson, P. E., & West, M. R. (2020). Results from the 2019 Education Next Survey. *Education Next*, *20*(1), 10–27.

Hill, H. C., & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Educational Research Journal*, *55*(5), 1076–1112. https://doi.org/10.3102/0002831218769614

Hinton, M. (2016, December 28). Arizona state school board votes to replace Common Core. *Education Week*. https://www.edweek.org/teaching-learning/arizona-state-school-board-votes-to-replace-common-core/2016/12

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R. C., & Staiger, D. O. (2016). *Teaching higher: Educators' perspectives on Common Core implementation*. Harvard University.

Kober, N., McIntosh, S., & Rentner, D. S. (2013). *Year 3 of implementing the common core state standards: Professional development for teachers and principals*. Center on Education Policy.

Kober, N., & Rentner, D. S. (2011). *States' progress and challenges in implementing Common Core State Standards*. Center on Education Policy.

Kober, N., & Rentner, D. S. (2012). *Year 2 of implementing the Common Core State Standards: States' progress and challenges*. Center on Education Policy.

Loveless, T. (2014). The 2014 *Brown Center report on American education: How well are American students learning? Part III: A progress report on the Common Core*. Brookings Institution.

Loveless, T. (2015). The 2015 *Brown Center report on American education: How well are American students learning? Part II: Measuring effects of the Common Core*. Brookings Institution.

Loveless, T. (2016). The 2016 *Brown Center report on American education: How well are American students learning? Part I: Reading and math in the Common Core era*. Brookings Institution.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for education reform*. U.S. Government Printing Office.

Norton, J., Ballinger, S., & Ash, J. (2016). *Massachusetts English language arts/literacy and mathematics curriculum frameworks review*. Abt Associates.

Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal*, *52*(6), 1185–1211. https://doi.org/10.3102/0002831215584435

Polikoff, M. S. (2017). Is Common Core "working?" And where does Common Core research go from here? *AERA Open*, *3*(1), 1–6. https://doi.org/10.1177/2332858417691749

Polikoff, M. S., Hardaway, T., Marsh, J. A., & Plank, D. N. (2016). Who is opposed to the Common Core and why? *Education Researcher*, *45*(3), 263–266. https://doi.org/10.3102/0013189X16651087

Rentner, D. S. (2013). *Year 3 of implementing the common core state standards: An overview of states' progress and challenges*. Center on Education Policy.

Rentner, D.S., & Kober, N. (2014). *Common core state standards in 2014: Districts' perceptions, progress, and challenges*. Center on Education Policy.

Schmidt, W. H., & Houang, R. T. (2012). Curricular coherence and the Common Core State Standards for mathematics. *Educational Researcher*, *41*(8), 294–308. https://doi.org/10.3102/0013189X12464517

Schmidt, W. H., McKnight, C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims, Volume I: A cross-national investigation of curricular intentions in school mathematics*. Kluwer.

Scholastic & the Bill & Melinda Gates Foundation. (2014). *Teachers' views on the Common Core State Standards one year later.* https://www.scholastic.com/primarysources/teachers-on-the-common-core.htm

Shadish, W. R., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Smith, M. S., & O'Day, J. A. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Falmer Press.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis*, *21*(1), 1–27. https://doi.org/10.3102/01623737021001001

St. Clair, T., Hallberg, K., & Cook, T. D. (2016). The validity and precision of the comparative interrupted time-series design: Three within-study comparisons. *Journal of Educational and Behavioral Statistics*, *41*(3), 269–299. https://doi.org/10.3102/1076998616636854

U.S. Department of Education. (2009). *Race to the top program executive summary*. https://files.eric.ed.gov/fulltext/ED557422.pdf

U.S. Department of Education. (2012). *ESEA flexibility (updated June 7, 2012)*. https://www2.ed.gov/policy/eseaflex/approved-requests/flexrequest.doc

What Works Clearinghouse. (2020). *What Works Clearinghouse procedures handbook (Version 4.1)*. Institute of Education Sciences, U.S. Department of Education.

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. https://doi.org/10.1080/19345747.2013.878011

Xu, Z., & Cepa, K. (2018). Getting college-ready during state transition toward the Common Core State Standards. *Teachers College Record*, *120*(6), 1–36. https://eric.ed.gov/?id=EJ1162833

Manuscript received January 18, 2020
Final revision received August 30, 2021
Accepted September 29, 2021