University of Chicago

BA Thesis

Race to the Top's Effects on Student Test Scores and High School Graduation Rates

Author: Mauro Ampie
Preceptor: Karlyn Gorski
Second Reader: Dr. David Johnson

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Arts

in

Public Policy Studies
The College

April 19, 2020

Abstract

Implemented from 2010 to 2015, Race to the Top was an application-based federal program that sought to increase states' usage of standards and accountability measures in education. Previous research has shown that states that won Race to the Top grants went on to increase their use of standards and accountability measures more than non-winning states, but no study has estimated the program's effects on student outcomes. In this thesis I fill this gap in the literature by using econometric analyses to calculate the program's effects on Math scores, Reading scores, and high school graduation rates. I find that states who won grants in 2010 improved their test scores by about 0.6 standard deviations, but that states who won grants in 2011 only improved their test scores by about 0.3 standard deviations. I also find that 2010 winners increased their high school graduation rates by 2%, whereas 2011 winners did not see improvements in this measure. Thus, in both test scores and high school graduation rates, Race to the Top had a larger positive impact on the 2010 winners than on the 2011 winners. The 2010 grants were about ten times as large as the 2011 grants, and I ultimately conclude that this difference in funding explains the disparity in gains between the 2010 winners and 2011 winners. Overall, my findings provide evidence that standards, accountability measures, and application-based programs can be used to improve student outcomes, and I recommend for the federal government to implement more standardization programs in the future.

Acknowledgements

I would like to extend my deepest thanks to the people that made this project possible. To my preceptor Karlyn Gorski, thank you for being available to discuss ideas and revise writing at every point of the process. To my second reader Dr. David Johnson, thank you for providing key feedback that helped me better understand the scope and limitations of my work. To Professor Anthony Fowler, thank you for helping me conduct the complex econometric analyses that form the backbone of this thesis. And to my friends Kyle Pinder and Mercedes Wentworth-Nice, thanks for always being willing to hear me talk about my thesis, no matter the time or day.

TABLE OF CONTENTS

Race to the Top: Effect on Student Outcomes

Since the passage of No Child Left Behind in 2002, standards and assessments have been at the forefront of federal education reform (Lohman, 2010). Building upon the trends of No Child Left Behind, in 2009 President Obama announced Race to the Top. The aim of Race to the Top was to increase the use of standards and accountability measures in the U.S. public education system. States competed against each other to receive Race to the Top grants, and funds were allocated to the states most willing to implement the guidelines and procedures outlines in the program application (The Obama White House, 2009). Previous research has shown that states awarded grants did increase their use of the practices and procedures prescribed by the program (Dragoset et al., 2016), and in this regard Race to the Top can be considered successful. However, the program also had significant shortcomings. Race to the Top required states to consider student scores when evaluating teachers and principals, and these score-based evaluations were then used to make high-stakes decisions (such as rehiring). This practice has been directly condemned by many researchers, who argue that evaluations based on student scores should only be used as feedback for improving instruction (Baker et al., 2013; Haertel, 2009). Furthermore, although no researcher has analyzed the program's effects on school climate and school staff morale, teachers and principals recount that Race to the Top had detrimental effects on these two measures (Winerip, 2011). Thus, the evidence presents a slightly mixed picture: Race to the Top led to increased use of standards and accountability measures, but it is unclear if these system-level improvements translated to concrete gains at the school and classroom level.

The conversation about Race to the Top's effectiveness has so far concentrated on the program's system-level effects, but to fully evaluate the program as effective or ineffective it is also necessary to look at the program’s effects on student outcomes. If the gains in standardization and accountability come at the expense of lower graduation rates and test scores, then perhaps one would no longer see the program as positive. Therefore, a crucial step in determining the program’s holistic impacts is to look at its effect on standardized test scores and graduation rates. Dragoset et al. (2016) attempts to examine the program’s effect on high school graduation rates and NAEP Math and Reading scores, but on both metrics the authors conclude that the evidence is insufficient to reach clear conclusions. More specifically, according to Dragoset et al. the graduation rate data is too incomplete for proper analysis, and the NAEP scores used (from 2003 to 2013) are insufficient to establish the pre-program test score trends. Dragoset et al. (2016) therefore conclude that the data on high school graduation rates and NAEP scores is insufficient to clearly establish Race to the Top’s effects on these measures.

To overcome the problem of insufficient data, in this thesis I look at test data from a longer range of years (1990 to 2015) and graduation rate data from two different measures (averaged freshman graduation rate and adjusted cohort graduation rate). Using these larger pools of data, I conduct differences-in-differences econometric analyses controlling for state fixed effects and year fixed effects. Through this methodology I calculate Race to the Top’s effects on test scores and high school graduation rates while accounting for the specific pre-program trajectories and any general state or time trends. I find that states who won grants in 2010 improved their test scores by about 0.6 standard deviations, but that the gains for the 2011 winners were only about half as large (0.3 standard deviations). Similarly, I find that 2010 winners increased their high school graduation rates by 2%, but that 2011 winners did not see improvements in this measure. Looking at possible explanations for why the 2010 and 2011 winners saw slightly different outcomes, I find that the 2011 winners likely experienced lower gains because the grants given out in 2011 were about ten times smaller than the grants given in 2010. In fact, the education budget of the 2011 winning states only increased by about 0.05% per year thanks to the Race to the Top grant, whereas the budgets of the 2010 winners increased by 1% per year thanks to the Race to the Top grants. Thus, I ultimately conclude that the small gains seen by the 2011 grant winners occurred thanks to the process of applying to the grants, such that these states saw improvements just from developing a careful and well-developed plan for increasing standards and accountability. Race to the Top therefore does seem to have led to small but important improvements in test scores and graduation rates. Overall, my findings provide evidence that standards and accountability measures can be used to improve student outcomes. And furthermore, my findings suggest that federal and state governments should implement more programs like Race to the Top in the future.

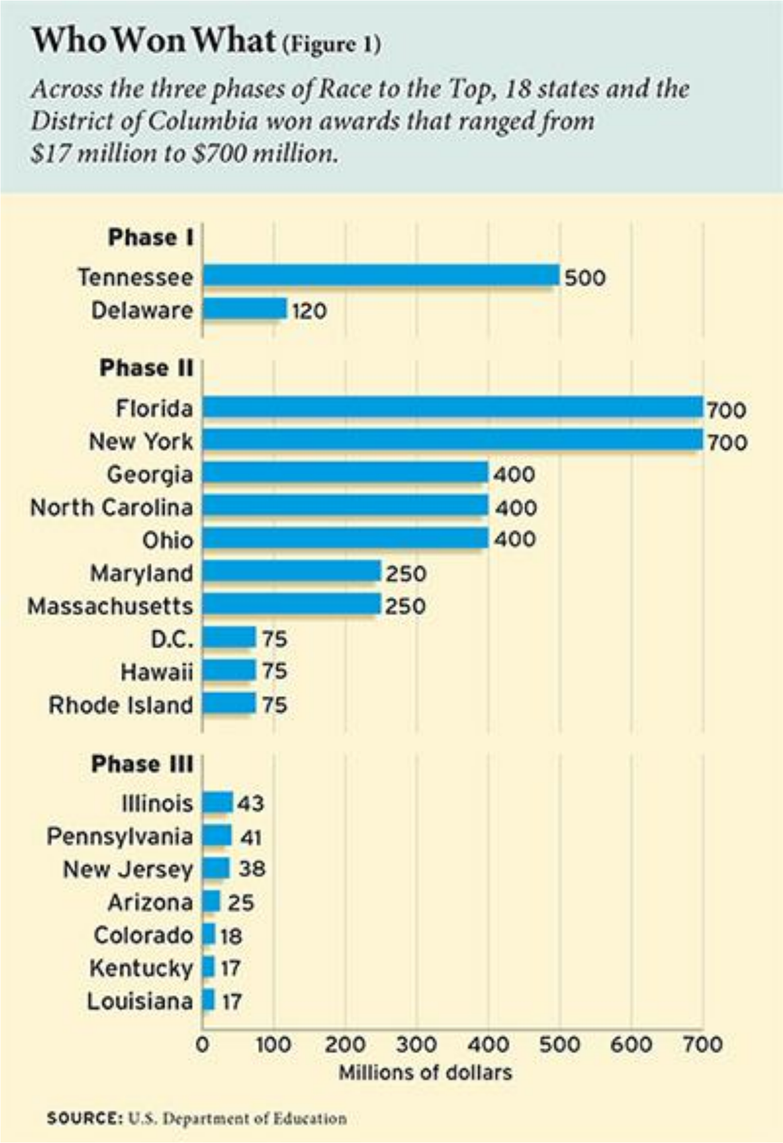
Program Design

In July of 2009 President Barack Obama announced Race to the Top, an education reform program that allocated $4.35 billion to states that agreed to increase their use of standards and assessments (The Obama White House, 2009). At the time of the program's inception Congress was largely unwilling to revise federal education law, so Race to the Top was an attempt by Obama and the U.S. Education Department to pursue educational reform at the state level (Tatter, 2015). The program was also created to help revitalize the states' education budgets that had recently shrunken as a result of the 2008 economic recession (Tatter, 2015).

Race to the Top was implemented parallel to No Child Left Behind, which was the main determinant of federal education policy from 2002 to 2015 (Klein, 2015). No Child Left Behind required states to develop and implement academic standards, make annual progress towards students' improvement goals, and reform struggling schools (Lohman, 2010). Race to the Top was therefore not the first program to implement standards and accountability measures, but it was certainly a pioneer in focusing exclusively on these system-level measures without directly tying them to goals on student outcomes.

Race to the Top only granted awards to states that agreed to implement the U.S. Education Department's reform agenda. Race to the Top had three application rounds (two rounds in 2010 and one round in 2011), and states that did not receive grants in previous rounds could modify their applications and reapply in later phases (Howell, 2015). Forty states and the District of Columbia applied for the first round, but only two states were given grants. The second round had ten winners and the third round had seven winners, but the third-round winners got grants about ten times smaller than the grants given to the winners of the 2010 application rounds. Regardless, by the end of all three applications rounds, eighteen states and Washington

D.C. had received Race to the Top funding (Lohman, 2010). The winning states and grant amounts for each round can be seen in the following table from Howell (2015):



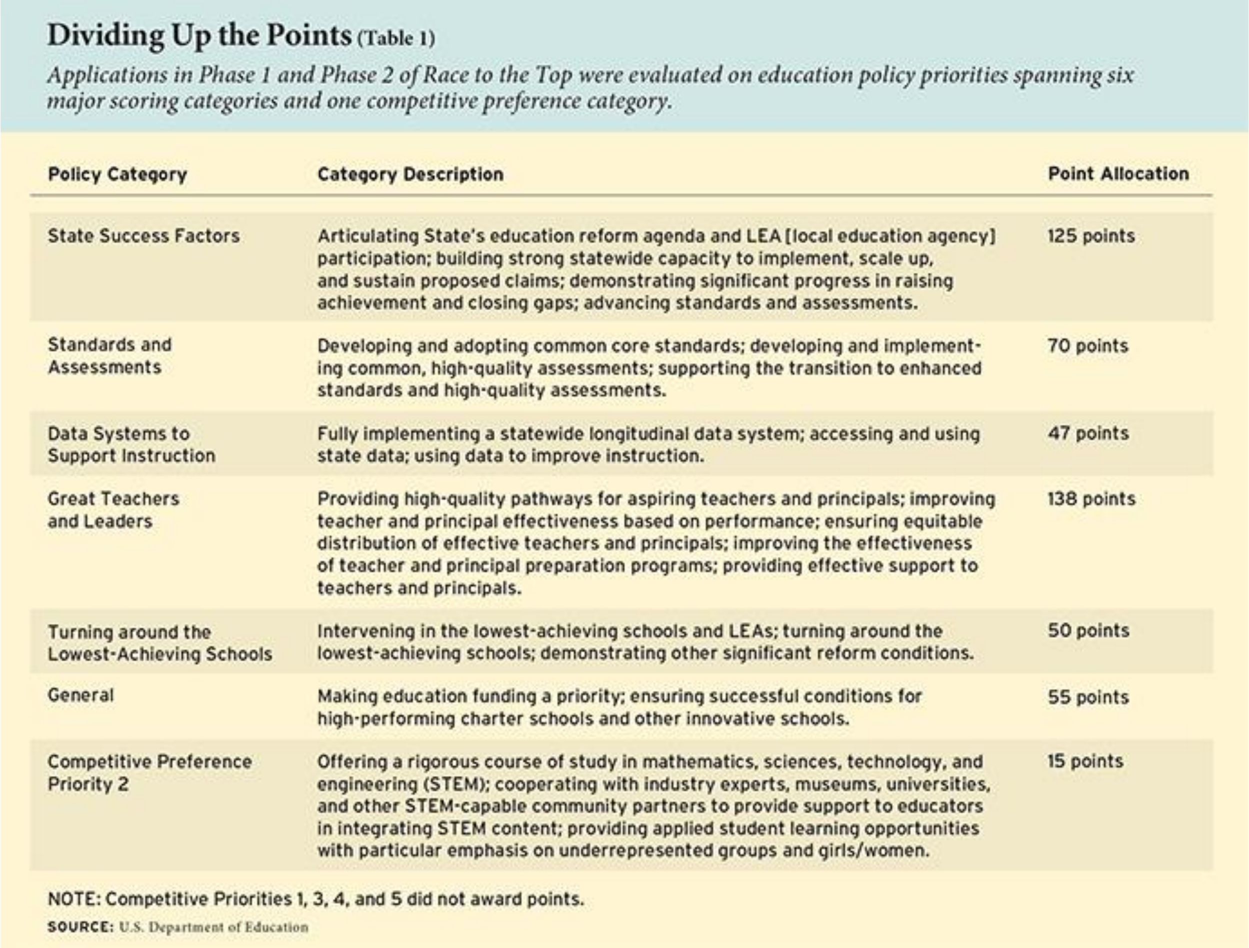
Who Won What (Figure 1)

Across the three phases of Race to the Top, 18 states and the District of Columbia won awards that ranged from $17 million to $700 million.

SOURCE: U.S. Department of Education

Once a state received a grant they had four years to use the grant money (Miller and Hanna, 2014), so the first-round and second-round winners had to use the funds by 2014 and the third-round winners had to use them by 2015.

Race to the Top only granted federal funds to states that agreed to implement the U.S. Education Department's reform agenda. The Education Department was thus responsible for establishing the operational and design guidelines that states would have to follow in their applications (Howell, 2015). The Education Department forced the states to pursue the following requirements in their Race to the Top applications: the creation of standards for the content and timing of student curriculum, the improvement of the lowest-performing schools, the measurement of student growth over time, and the design of policies to reward and retain the best teachers (Howell, 2015; U.S. Department of Education, 2016). The student growth metrics and teacher evaluations were combined into a single system under which teacher performance would be directly linked to student performance. The specific criteria used to evaluate the states' Race to the Top applications can be seen in the following table from Howell (2015):



Dividing Up the Points (Table 1)

Applications in Phase 1 and Phase 2 of Race to the Top were evaluated on education policy priorities spanning six major scoring categories and one competitive preference category.

Policy Category	Category Description	Point Allocation
State Success Factors	Articulating State's education reform agenda and LEA [local education agency] participation; building strong statewide capacity to implement, scale up, and sustain proposed claims; demonstrating significant progress in raising achievement and closing gaps; advancing standards and assessments.	125 points
Standards and Assessments	Developing and adopting common core standards; developing and implementing common, high-quality assessments; supporting the transition to enhanced standards and high-quality assessments.	70 points
Data Systems to Support Instruction	Fully implementing a statewide longitudinal data system; accessing and using state data; using data to improve instruction.	47 points
Great Teachers and Leaders	Providing high-quality pathways for aspiring teachers and principals; improving teacher and principal effectiveness based on performance; ensuring equitable distribution of effective teachers and principals; improving the effectiveness of teacher and principal preparation programs; providing effective support to teachers and principals.	138 points
Turning around the Lowest-Achieving Schools	Intervening in the lowest-achieving schools and LEAs; turning around the lowest-achieving schools; demonstrating other significant reform conditions.	50 points
General	Making education funding a priority; ensuring successful conditions for high-performing charter schools and other innovative schools.	55 points
Competitive Preference Priority 2	Offering a rigorous course of study in mathematics, sciences, technology, and engineering (STEM); cooperating with industry experts, museums, universities, and other STEM-capable community partners to provide support to educators in integrating STEM content; providing applied student learning opportunities with particular emphasis on underrepresented groups and girls/women.	15 points

NOTE: Competitive Priorities 1, 3, 4, and 5 did not award points.

SOURCE: U.S. Department of Education

All seven categories (turning around lowest performing schools, rewarding good teachers, establishing learning standards, etc.) can ultimately be condensed to two broader themes: establishing specific benchmarks for student achievement, and holding teachers and schools accountable for the performance of their students. Thus, despite the many distinct categories in the application guidelines, Race to the Top can simply be conceptualized as a program for increasing the use of standards and accountability measures within the educational system. Systems, rather than outcomes, were the core focus of the program.

Theoretical Framework: The Promises and Pitfalls of Standardization

With No Child Left Behind and Race to the Top the educational field has recently started to pursue standards and accountability, but the medical field has championed standardization for well over half a century (Timmermans and Berg, 2010). As such, looking at standardization in medicine can help us understand the benefits and challenges of standardizing education.

In medicine, standardization consists of using scientific evidence to determine guidelines for clinical practice (Timmermans and Berg, 2010; Bothwell et al., 2016). Under ideal circumstances medical guidelines are developed from scientific evidence collected throughout several studies, but oftentimes guidelines are instead developed by expert panels that rely on their professional experiences more than on scientific evidence (Timmermans and Berg, 2010; Bothwell et al., 2016). The picture is similar in the educational field: best practices are defined as practices that possess widely-agreed effectiveness, combining professional wisdom with empirical evidence (Alber, 2015; Moran and Malott, 2004). However, just like in medicine, educational standards are often based on professional wisdom rather than on empirical evidence (Common Core State Standards Initiative, 2019).

One of the main arguments in favor of standardization is that it can increase efficiency. By carefully examining the benefits and drawbacks of specific practices and interventions, the most cost-effective and optimal can be determined and then widely implemented throughout the field. In medicine, standardization has often led to better and cheaper clinical practices, which in turn has led to improved patient outcomes (Timmermans and Berg, 2010; Bothwell et al., 2016). For example, extensive research on asthma has led to the creation of clinical guidelines for the treatment of asthmatic patients, and these guidelines have been shown to be less expensive and lead to better patient outcomes than alternative treatment methods (Timmermans and Mauck, 2005). In the United States., teaching tends to occur behind closed doors, with teachers having almost complete control over what occurs in their classroom (Ripley, 2013). Ensuring the usage of best practices is therefore likely more challenging in the educational system than in the medical sphere. However, with programs like Race to the Top and No Child Left Behind the educational field is certainly pushing toward a wider usage of best practices, and based on the results seen in medicine this push should eventually help improve the quality of education received by children all throughout the United States.

Another important argument for standardization is that it increases transparency. All service providers are evaluated along the same guidelines, and stakeholders can access the publicly available guidelines of best practice to ensure that the service providers (be it teachers or doctors) are conducting work backed by research findings. Thanks to standardization, patients can now ensure that they are receiving treatment in accordance with medical clinical guidelines (Timmermans and Berg, 2010; Gabbay and Le May, 2010). Similarly, standardization allows school staff and parents to know the exact criteria by which students and schools are evaluated, making evaluation processes more transparent (Bothwell et al., 2016, Greenhalgh et al., 2014; Howell, 2015). Furthermore, thanks to clear evidence on what constitutes effective teaching, principals can follow standard observation procedures to provide detailed feedback to teachers (Ing, 2010). Teachers can then use this feedback to improve the quality of their instruction. In summary, standardization has been used for decades in the medical field to increase efficiency and transparency, and based on these results in the medical field, standardizing education seems like a promising way to make schools and districts more transparent and effective.

Despite the advantages that standardization has brought to the medical field, two problems have curtailed standardization’s capacity to effect meaningful reform in medicine.

First, in imposing guidelines for best practice, standardization has limited the jurisdiction of doctors and has forced them to align with the standard practices even when they deem other practices necessary (Timmermans and Berg, 2010; Bothwell et al., 2016; Timmermans and Mauck, 2005; Cook, 2015). Guidelines for best practice often fail to take into account individual characteristics of patients, which makes the guidelines ineffective under certain circumstances (Timmermans and Berg, 2010; Greenhalgh et al., 2014; Cook, 2015). Thus, overly prescriptive imposition of guidelines has made physicians unable to use these guidelines in the most effective manner. The second problem with standardization is that the guidelines for best practice are often determined by outside actors with little on-the-ground experience (Bothwell et al., 2016, Greenhalgh et al., 2014; Howell, 2015). These outside actors (insurers, government agencies, etc.) cannot properly envision the practical difficulties of implementing specific guidelines, so established guidelines are often impractical to implement (Haertel et al., 2009; Bothwell et al., 2016, Greenhalgh et al., 2014; Howell, 2015). If the guidelines were more readily usable, doctors would have greater capacity to implement them.

Standardization is a recent newcomer to the educational field, but it has shown itself mainly in two forms: the creation of consistent learning standards, and the widespread implementation of teacher and school accountability measures. No Child Left Behind and Race to the Top have both been important drivers of this standardization process. Under these programs, states began to implement the same procedures for measuring student learning and evaluating teachers and schools (Lohman, 2010). And despite standardization just recently entering the educational arena, researchers have already started to examine its effects on teacher instruction and student outcomes.

Student test scores is one metric that does seem to improve as a result of using standards and accountability measures. A meta-analysis of 14 studies finds that test-driven accountability policies slightly increase student scores on Math and Reading (Lee, 2008). Hanushek and Raymond (2005) also find a positive effect, and they show that accountability systems implemented in the 1990s led to improvements in student achievement scores. An important caveat, however, is that the researchers only find improvement in the states that attached consequences to performance measures. Another study finds an increase in the scores of low-performing students from low-achievement schools as a result of the schools' increased implementation of accountability measures (Deming et al., 2016). Lastly, research has shown that No Child Left Behind (which focused on standards and accountability) did not have an effect on fourth-grade Reading achievement but did increase the average Math performance of fourth graders and eighth graders, particularly among traditionally low-achieving groups (Dee and Jacob, 2011). Thus, the research evidence generally suggests that implementing standards, assessments, and accountability measures leads to increases in student scores.

For other student outcomes, however, standards and accountability do not seem beneficial. One study finds that implementing extensive accountability in high-achieving schools led to a decrease in high school graduation rates for the schools' low-performing students, as well as a decrease in their yearly salary at age 25 (Deming et al., 2016). Similarly, Carnoy and Loeb (2002) find that states with high accountability measures did not have lower student grade retention or high school dropout rates, even though students in these states saw increases in their test scores. In summary, for student outcomes other than test scores (grade retention, graduation rates, and post-graduation salary), increased use of standards and accountability measures has either a neutral or negative effect. These set of findings are problematic because, other than improving student scores, standards and accountability measures do not seem to have a positive effect on students' educational careers.

Besides its negative impacts on students, standardization sometimes also negatively affects teachers. Under Race to the Top, states widely adopted the value-added teacher evaluation system This system catalogues teachers as effective or ineffective based on their students' performance on standardized tests (Lohman, 2010). These evaluations are then used to determine if teachers are rehired or terminated. Unfortunately, using the value added-system in this way is ill-founded. The value-added evaluation system does not have the statistical reliability and validity to be used in high-stakes decisions regarding teachers because student scores fluctuate significantly from year to year (Baker et al., 2013; Hill, 2009; Haertel, 2009). Additionally, the value-added evaluation system does not seem to improve teacher quality or student learning. Programs that implement the value-added evaluation system do not see improvements in students' standardized test scores (Gallivan, 2019), and teacher instruction is actually worsened under the value-added system because teachers begin teaching to the test and make their instruction more procedural and less conceptual (Olah et al., 2010). Perhaps most problematic, the widespread usage of the value-added system seems to have lowered teacher morale and job satisfaction because teachers feel like they are being evaluated based on something largely out of their control (Winerip, 2011). Standardizing education has led to widespread adoption of the value-added teacher evaluation system, but in this way standardization has negatively impacted teachers' satisfaction and instructional practices.

Race to the Top: Previous Research

As mentioned in the previous section, standardization in education has often failed to improve student outcomes and teacher performance. And in this regard the evidence does not bid well for Race to the Top, seeing as the program explicitly pushed for practices that in the past have not led to improvements for students or teachers. However, even if the program's components were based on little evidence of past effectiveness, Race to the Top certainly succeeded in getting states to adopt the program's measures. Race to the Top sought to increase states' usage of standardized learning goals, and the winning states did end up implementing more similar learning curricula (Dragoset et al., 2016). Likewise, under Race to the Top states widely adopted the value-added teacher evaluation system, a system that policymakers lauded for rewarding good teachers and holding bad teachers accountable (Dragoset et al., 2016). The grant-winning states also went on to institute a large number of policies related to standards and accountability measures through their state legislatures, which suggests that the program succeeded in spurring state-led standardization initiatives (Howell and Magazinnik, 2017). And the state legislatures of non-winning states also passed a fair number of standardization policies during the Race to the Top years, which suggests that the program spurred standardization even among non-winning states (Howell and Magazinnik, 2017).

Previous evidence suggests that standards and accountability measures often fail to improve student outcomes and teacher performance (Lee, 2008; Hanushek and Raymond, 2005; Carnoy and Loeb, 2002; Deming et al., 2016), but Race to the Top' effects on these metrics has been left largely unexplored. The program did not explicitly mention improvement of student outcomes or teacher performance as one of its goals, so it is perhaps natural that up to now the research on Race to the Top has focused mostly on other metrics. The program's effect on instructional practices has been left completely unexamined, and the few attempts to quantify the program’s effects on student outcomes have been ultimately inconclusive. Dragoset et al. (2016) try to evaluate Race to the Top’s effect on student scores and high school graduation rates, but for both measures the researchers conclude that the data is insufficient to arrive at reliable conclusions. The researchers do not attempt any analyses on high school graduation rates because the graduation rate data they acquire is not consistent enough between states. Dragoset et al. do have test score data consistent across states and conduct extensive analyses with this data, but they eventually refrain from publishing their findings due to concerns about what type of trajectory the student scores were following prior to Race to the Top. Specifically, Dragoset et al. conclude that the available data (NAEP Math and Reading scores from 2003 to 2015) is insufficient to reliably determine if the trend seen prior to 2009 was linear or exponential. Dragoset et al. instead resort to a purely descriptive analysis of the data trends, but the authors ultimately deem this methodology insufficient as well. Thus, despite significant efforts, Dragoset et al. (2016) are unable to confidently estimate Race to the Top’s effects on student scores.

To fill in this gap in the Race to the Top literature, in this thesis I use high school graduation rate data from two different measures (AFGR and ACGR) and test data from two subjects (Math and Reading) to determine Race to the Top’s effects on graduation rates and student scores. Crucially, I overcome Dragoset et al.’s problem of insufficient data by looking at a larger range of test score years (1990 to 2015) and by looking at two measures of graduation rates that, once combined, also provide consistent data across states from 1990 to 2015.

Data Sources

To examine Race to the Top's effects on student outcomes, in this thesis I utilize two publicly available measures: NAEP student test scores and high school graduation rates.

NAEP Test Scores:

NAEP tests are standardized assessments conducted at 4th, 8th, and 12th grade, although the range of subjects tested at each grade level varies slightly. I try looking at the student scores for all exams administered by NAEP, but only nine have results available at the state level. And of those nine exams, only four have been administered for enough years to support the assumptions of my differences-in-differences econometric analyses. Therefore, in this thesis I ultimately only end up analyzing scores for the four tests with enough data: Math 4th grade, Math 8th grade, Reading 4th grade, and Reading 8th grade. The Math assessments were first implemented in 1990 and the Reading assessments were first implemented in 1992, so the year range I examine is 1990-2015 for Math scores and 1992-2015 for Reading scores.

Two technical details about NAEP scores are worth mentioning. First, NAEP assessments are administered to a representative sample of students, so NAEP scores are an accurate picture of national student achievement (NAEP, 2018). And second, NAEP provides group-aggregated data but does not provide individual-level data. So, for example, one can access the average performance of Alabama students in 2007 but cannot access the 2007 score of each Alabama student tested. Fortunately, I do not need individual-level for my analyses because I only conduct comparisons at the state level. And although using aggregated data instead of individual data can sometimes slightly affect results, the impacts are almost always tiny and do not change the overall result patterns (Robinson, 1950; Jacob, 2016).

The NAEP test score data has three key limitations:

(1) Most exams administered only have data at the national level, meaning that I can only conduct analyses on the exams that do provide state-level data. Even then, assessments are only conducted on 4th, 8th and 12th grade. Thus, the conclusions I derive from my analyses apply only to the subjects and grades for which data is available. It seems reasonable to assume that Race to the Top's effects on these specific subjects and grades are representative of the program's more generalized effects, but only an analysis of trends in more grades and subjects could substantially answer this question. Thus, the NAEP scores provide a useful but partial picture of Race to the Top's effects on student scores.

(2) Some researchers claim that the characteristics of students tested in each state varied from year to year, which would make any analysis incapable of disentangling the program effects from the effects of changing the sample composition (Dragoset et al., 2016). NAEP ensures that the students tested are representative of the national student population (NAEP, 2018), but representation at the national level does not necessarily translate to representativeness within each state. The problem of changing sample composition could be addressed by looking at scores broken down by gender, race, and socioeconomic status (which NAEP does provide), but states do not always report test score data by demographics. Therefore, the test score data broken down by demographics is much more incomplete than the aggregated data. I ultimately only conduct analyses on the aggregated data so that I could accurately reflect students' performance in all states, but this methodological choice limits my capacity to control for effects caused by demographic changes in each state. Since my analyses are conducted at the group level (2010 grant winners, 2011 grant winners, and non-winners) the demographic changes would only affect my results if they consistently occur only in one specific group and not in the

others. Regardless, since my methodological choices make me unable to closely control for demographic changes, future research should look to replicate my analyses with the data disaggregated by race.

(3) The goal of NAEP is to survey the knowledge of students across the nation in a broad range of content and skills, so it purposely does not align with any specific state curriculum. It could therefore be the case that, if a state's curriculum and the NAEP test topics do not coincide, a state could be improving their students' understanding of the state curriculum despite no effect shown in the NAEP scores. This problem is probably of little relevance (the topics covered by the state curriculum and the NAEP tests are likely similar), but states could certainly argue that the NAEP scores fail to reflect the full improvements taking place at the state level (Haertel, 2009). However, because scores must come from a single examination to allow for between-state comparisons, NAEP scores are the best measure available at present for evaluating nationwide education programs.

In summary, the NAEP data is not perfect but still provides a fairly accurate and representative picture of students' academic performance through the years.

High School Graduation Rates:

For high school graduation rates I look at two measures: averaged freshman graduation rates (AFGR) and adjusted cohort graduation rates (ACGR). Both AFGR and ACGR measure the percentage of students that graduate with a regular diploma four years after entering ninth grade. However, AFGR estimates the size of the entering ninth grade class from aggregated student enrollment data, whereas ACGR estimates the size of the class more accurately by removing students that die or transfer out of the state during high school (NCES, 2020). ACGR is therefore more accurate representation of graduation rates, but states have only recently started to collect and release the data needed to calculate ACGR (NCES, 2020). As such, AFGR is available from 1960 to 2013, whereas ACGR is only available starting from 2011. For this thesis I unite AFGR and ACGR into a single combined measure, which allows me to track high school graduation rates continuously from 1990 to 2015. And while not calculated in exactly the same way, AFGR and ACGR are similar enough to be generally comparable. For example, I compare AFGR and ACGR from 2011 to 2013 (the years for which both measures are concurrently available) and confirm that the standard deviations and averages of both measures are very similar and typically only differ by a couple percentage points (see Appendix A). Furthermore, I confirm that both measures show a slight increase in graduation rates every year from 2011 to 2013, which suggests that they are capturing the same nationwide trends. I even look at group-specific trends (2010 winners, 2011 winners, and non-winners) and confirm that within these groups AFGR and ACGR also behave in the same way, which eliminates any potential within-group difference between the two measures that could have remained hidden during aggregated analyses. For a more detailed explanation of my analyses comparing AFGR and ACGR, see Appendix A. In short, through careful examination of the two measures I confirm that they are comparable enough to be combined into a single measure for my econometric analyses.

The high school graduation rate data has two key limitations:

(1) ACGR is measured more consistently among states than AFGR, which means that direct comparisons between states are less reliable with AFGR than with ACGR. The majority of graduation rate data I use comes from AFGR, seeing as AFGR is the only high school graduation rate measure available before 2011. Thus, my analyses on graduation rates would be severely compromised if AFGR is not a usable measure. My analyses readily adjust for between-state

differences in the formulas used to calculate AFGR, but my analyses would be strongly compromised if states changed their AFGR calculation formulas from year to year. For example, if Alabama changed their calculation formulas from year to year, the trends in Alabama graduation rates could be caused by changes within the student body or changes within the formulas themselves. As such, my analyses only stay valid if states used the same calculation formulas through the years. Fortunately, the NCES documentation on AFGR only mentions that AFGR suffers from between-state consistency, so it seems that AFGR does have within-state consistency (NCES, 2020). Thus, AFGR ultimately does seem like a valid measure for my analyses.

(2) With the rise of school accountability in the last two decades, some researchers worry that schools have begun to artificially inflate high school graduation rates in order to meet federal requirements (Harris et al., 2020). Thus, any increases seen across the nation could be reflective of bureaucratic changes on calculation formulas rather than substantial improvements. Some anecdotal evidence does suggest that states have tampered with their calculations of high school graduation rates to create artificial improvements, but for the most part the increases seem to occur as a result of substantial improvements in students' learning capital (Harris et al., 2020). And even if states engaged in strategic behavior to artificially improve their high school graduation rates, as long as both Race to the Top winners and non-winners engaged in this behavior, my statistical models account for these artificial increases.

In summary, despite some limitations, AFGR and ACGR are both valid measures with which to conduct my econometric analyses.

Methods

To calculate Race to the Top's effects on student test scores and high school graduation rates, I use a unit fixed effects regression model:

$$\begin{aligned} \mathbf{DV} = {} & \boldsymbol{\beta}\mathbf{1 * \text{ReceiveRTT}} + \boldsymbol{\beta}\mathbf{2 * \text{dummyAlabama}} + \boldsymbol{\beta}\mathbf{3 * \text{dummyAlaska}} + \boldsymbol{\beta}\mathbf{3 * \text{Arizona}} + \ldots \\ & + \boldsymbol{\beta}\mathbf{51 * \text{Wisconsin}} + \boldsymbol{\beta}\mathbf{52 * \text{Wisconsin}} + \boldsymbol{\beta}\mathbf{53 * \text{dummy1990}} + \boldsymbol{\beta}\mathbf{54 * \text{dummy1991}} + \ldots \\ & + \boldsymbol{\beta}\mathbf{77 * \text{dummy2014}} + \boldsymbol{\beta}\mathbf{78 * \text{dummy2015}} \end{aligned}$$

The model is mathematically complex, but conceptually very simple. DV is the dependent variable (test scores or high school graduation rates). ReceiveRTT is a dummy variable with value 1 when a state has a Race to the Top grant in that year and 0 when a state does not have a Race to the Top grant in that year. DummyAlabama, along with all the other state dummy variables, takes a value of 1 for that specific state (in this case Alabama) and 0 for all other states. Similarly, dummy 1990, along with all other year dummy variables, takes a value of 1 for that specific year (1990) and a value of 0 for all other years. In other words, the model calculates the effect of receiving a Race to the Top grant while controlling for state effects (a state's pre-Race to the Top performance) and year effects (nationwide decreases or increases in the dependent variable).

The unit fixed effects regression model is a modified version of the standard differences-in-differences econometric model. In the context of estimating Race to the Top's effects, the standard differences-in-differences econometric model would look as follows:

$$\mathbf{DV} = \boldsymbol{\beta}\mathbf{1 * \text{RTT}} + \boldsymbol{\beta}\mathbf{2 * \text{TreatmentYear}} + \boldsymbol{\beta}\mathbf{3 * \text{ReceiveRTT}}$$

In this model, RTT is a binary variable with value 1 for Race to the Top winning states and 0 for all Race to the Top non-winning states, regardless of year. TreatmentYear is also a binary variable, taking value 1 for years in which Race to the Top is implemented and 0 for years in which Race to the Top is not implemented. And like in the unit fixed effects regression model, ReceiveRTT is a dummy variable with value 1 when a state has a Race to the Top grant in that year and 0 when a state does not have a Race to the Top grant in that year. In other words, in the standard differences-in-differences model RTT and TreatmentYears are the main effect variables and ReceiveRTT is the interaction variable. And just like the unit fixed effects model, the standard differences-in-differences model calculates Race to the Top's effects on student test scores and high school graduation rates. Ultimately, however, the unit fixed effects model is more appropriate because it is more precise than the standard differences-in-differences model. Thanks to including an independent dummy variable for every state and year, the unit fixed effects regression model can account for every single state's previous trajectory and every single year's effect. In short, while the unit fixed effects regression model and the standard differences-in-differences model both account for state and year effects, the unit fixed effects regression model accounts for these effects much more precisely. And because of this, social scientists tend to prefer this model over the standard differences-in-differences model, even though both models can ultimately be used to validly establish causal relationships (Somers et al., 2013; St. Clair et al., 2014; Jacob et al., 2016; Imai and Kim, 2019).

Controlling properly for state and year effects is crucial for validly determining Race to the Top's effects. As can be seen in Appendix B, both the grant-winning and non-grant-winning states were showing an increase in the student scores and high school graduation rates even before Race to the Top, and the Race to the Top grant winners were increasing their test scores more than the Race to the Top non-winners even before the program was implemented. This pattern of increases is the reason why previous attempts at estimating Race to the Top's effects on student outcomes have fallen short. Without enough data to reliably determine the pre-program trends, statistical models cannot ultimately determine if the post-program increases are caused by the program or are simply continuations of the pre-program increases. Fortunately, by using data from 1990 to 2015 I properly determine the pre-Race to the Top trends. Research suggests that for yearly data one needs at least four pre-treatment years to reliably establish baseline trends (Somers et al., 2013), so the twenty-year range before Race to the Top's implementation is more than enough to firmly establish both the size and trajectory of pre-Race to the Top trends. And thanks to this, the unit fixed effects regression model leads to correct calculations of Race to the Top's causal effect on test scores and high school graduation rates.

One important assumption behind the unit fixed effects model I use is that test scores and graduation rates increase or decrease linearly through the years. If for some reason test scores and graduation rates instead increase in some other way (for example, exponentially), then the model would misestimate the effects of receiving a Race to the Top grant. However, as can be seen in Appendix B, from 1990 to 2015 both the test scores and graduation rates increase by a relatively consistent amount every year. Therefore, the year-to-year increases in NAEP test scores and high school graduation rates are much better represented by a linear model than by any other model. And because of this, the unit fixed effects regression model I use is the most accurate model for determining Race to the Top's effects on student outcomes.

An important methodological choice I make is to separate the Race to the Top winners into two separate treatment groups: the 2010 winners, and the 2011 winners. The reason for this distinction is that the 2010 winners received grants about ten times as large, on average, as the grants received by the 2011 winners. Thus, the program’s effects might look different for the two groups. Furthermore, by separating the winners into two, I can more accurately account for the years in which Race to the Top is implemented. Thus, for the 2010 winners I consider 2011-2014 to be the treatment years, whereas for the 2011 winners I consider the treatment years to be 2012-2015. In this way, I evaluate the states’ performance during the four-year period in which they implemented Race to the Top.

Findings and Discussion

Implementing the unit fixed effects regression model on Math 4th grade, Math 8th grade, Reading 4th grade, and Reading 8th grade test scores, I find that Race to the Top had the following effects:

- In Math 4th grade scores, receiving a Race to the Top grant in 2010 led to an increase of 3.6 points (p = 0.0001) and receiving a Race to the Top grant in 2011 led to an increase of 1.7 points (p = 0.03)
- In Math 8th grade scores, receiving a Race to the Top grant in 2010 led to an increase of 4.4 points (p = 0.0001) and receiving a Race to the Top grant in 2011 led to an increase of 2.7 points (p = 0.004)
- In Reading 4th grade scores, receiving a Race to the Top grant in 2010 led to an increase of 5.1 points (p = 0.0001) and receiving a Race to the Top grant in 2011 led to an increase of 2.5 points (p = 0.0031)
- In Reading 8th grade scores, receiving a Race to the Top grant in 2010 led to an increase of 2 points (p = 0.0005) and receiving a Race to the Top grant in 2011 did not have an effect (p = 0.21).

One noteworthy aspect about this pattern of results is that in almost all cases (7 out of 8) receiving a Race to the Top grant translated into statistically significant gains in test scores. Additionally, the increases were always about twice as large for the 2010 winners as for the 2011 winners. And to put the point gains into perspective, within any given year each group (2010 winners, 2011 winners, and non-winners) had a spread anywhere between 20 to 40 points, meaning that in any given year the lowest performing state in a group performed about 20 to 40 points below the highest performing state in that same group. With such a relatively short range, an increase of 2 to 5 points does seem of considerable size. And in terms of standard deviations, the 4 to 5 point gains seen by 2010 winners and the 2 point gains seen by 2011 winners are increases of about 0.6 standard deviations and 0.3 standard deviations, respectively. Increases in standard deviations are still hard to conceptualize, but the literature generally suggests that an increase of 0.08 standard deviations is equivalent to about one extra month of schooling (Greenstone et al., 2012). Thus, based on my analysis, the students in 2010-winning states saw improvements equivalent to about seven extra months of schooling. Similarly, the 2011-winners saw increases of about four extra months of schooling. Put another way, for every year that the 2010 winners had the RTT grant, their students had gains equivalent to what they would have had if they had received about two extra months of schooling during that school year. Likewise, for every year that the 2011 winners had the RTT grant, their students had gains equivalent to what they would have had if they had received about one extra month of schooling during that school year. Thus, the pattern is clear: both 2010 winners and 2011 winners see increases in their test scores thanks to the Race to the Top grants, and the increases are about twice as large for the 2010 winners as for the 2011 winners.

The results are somewhat similar for high school graduation rates. Implementing the unit fixed effects regression model on my combined measure for high school graduation rates (composed of 1990-2013 AFGR data and 2014-2015 ACGR data), I find that Race to the Top had the following effects: receiving a grant in 2010 translated to a 2% increase in graduation rates ($p = 0.0004$), but receiving a grant in 2011 did not have any effect ($p=0.71$). Thus, unlike for test scores, in terms of graduation rates the Race to the Top grants only had a positive effect for the 2010 winners.

Overall, the results of the test score and graduation rate data show that the 2010 winners benefitted more from the program than the 2011 winners. This pattern of results might initially seem puzzling. Both groups implemented the same program, so in theory the two groups should have derived the same benefits from the program. Two explanations can account for this disparity in results between the two groups:

1) All the 2011 winners had applied for the first application rounds in 2010 but had not gained a grant in those rounds. Thus, it could be the case that the 2010 winners were the states most ready and willing to implement the program components of Race to the Top, which would suggest that they probably implemented the program to a larger degree. And if the two groups indeed differed in eagerness and capacity to implement the program, then the disparity in results could be caused by a between-group difference.

2) Alternatively, the disparity in results between the two groups could be caused by the difference in grant size. The 2010 grant winners received about 300 million dollars on average, whereas the 2011 grant winners received only about 30 million dollars on average. In other words, the 2011 grants were 10 times smaller than the 2010 grants. This disparity in funding could be compromising the 2011 winners' capacity to implement the plans they outlined in their applications, which would in turn mean that the program components of Race to the Top got implemented to a lesser degree in these states.

In theory both explanations are plausible, but the "between-group differences" explanation does not seem to align with evidence. The 2011 winners had applied for the two 2010 application cycles and had been rejected, but in the 2010 applications the 2011 winners received scores very close to the 2010 winners. Specifically, the 2010 winners had application scores ranging from 440 to 470, but the would-be 2011 winners were very close behind with scores between 410 and 440 (U.S. Department of Education, 2016). The 2011 winners did not receive an application score in the 2011 round, but since they were already scoring about as well as the 2010 winners in the 2010 cycle, their 2011 applications were likely as good as the 2010 applications of the 2010 winners. As such, it seems likely that upon receiving the grant money both groups had about equally good plans and capacity to implement Race to the Top. Thus, it seems highly unlikely that the disparate pattern of results between the 2010 and 2011 winners is caused by a substantial between-group difference in preparedness or eagerness to implement Race to the Top.

In contrast, evidence supports the explanation that the disparity in grant sizes caused the difference in results between the 2010 and 2011 winners. The 2010 winners received larger grant sizes, and the grants had a much larger impact on the yearly education budgets for the 2010 winners than for 2011 winners. Specifically, for the 2010 winners receiving a Race to the Top grant translated in average to a 1% increase in their education budget for the next four years, whereas for the 2011 winners receiving a Race to the Top grant translated in average to a meager 0.05% increase in their education budget for the next four years. I calculated these increases through the following formula:

Percent change in state ed budget = (Race to the Top Grant / 4) / 2010 state ed budget

The 4 in the above formula accounts for the fact that states had four years to use their grant money, such that the grant money would be about evenly split between the next four years. The formula is not a perfect estimate of the yearly grant impact because the 2010 budget is not a completely accurate reflection of the 2011-2014 budgets, but as long as states mostly maintained budget consistency from year to year, the formula accurately reflects the percentage by which a state's educational budget increased thanks to the Race to the Top grant money. And based on this formula, I find that the impact of Race to the Top grants is twenty times larger on the education budgets of the 2010 winners than on the education budgets of the 2011 winners (1% vs. 0.05%). This difference in impact is even larger than the tenfold difference one would predict from the dollar-for-dollar grant comparisons (average 300 million dollar grants for 2010 winners vs. average 30 million dollar grants for 2011 winners). Thus, the difference in grant size likely caused the disparity in results between the 2010 and 2011 winners.

The meager impact that Race to the Top grants had on the education budgets of the 2011 winners naturally begs the question: if the state budgets of the 2011 winners increased only by a meager yearly 0.05% thanks to Race to the Top, how did these states see any gains whatsoever? 0.05% of a state budget is still millions of dollars, but when divided by the number of children served in every state, the Race to the Top money only ended up being a couple extra dollars per student for the 2011 winners. So it is perhaps not surprising that the 2011 winners did not see gains for Reading 8 test scores and high school graduation rates. However, these states did see small but important gains in Math 4, Math 8, and Reading 4 scores. How did these gains occur if the grant money was unsubstantial for the 2011 winners? Two explanations come to mind:

1) Perhaps another educational program was implemented parallel to Race to the Top, and the gains seen by the 2011 winners occurred thanks to the other program rather than thanks to Race to the Top.

2) Perhaps the process of applying for a Race to the Top grant was in and of itself beneficial for the states. Maybe the development of a careful implementation plan for standards and accountability measures translated to gains despite no extra funding.

The first explanation initially sounds plausible, but it falls short on two fronts. First, for the explanation to be correct, the 2011 winners would need to be benefitting from the other program more than the controls. Since my unit fixed effects regression analysis is a comparison between Race to the Top winners and non-winners, any general education program that benefitted all states equally would show no difference between the performance of 2011 winners and non-winners. Thus, another educational program would only explain my findings for the 2011 winners if it consistently benefitted the 2011 winners more than the controls. Some inherent characteristic of the 2011 winners could theoretically have made them benefit more from a program implemented in all states, but no characteristic readily comes to mind. And second, Race to the Top was the main federal education program implemented throughout these years that awarded billions of dollars to state education budgets (Sass, 2020). Race to the Top happened parallelly to No Child Left Behind, but No Child Left Behind had been in place since 2002 so it would not account for any effects seen immediately at 2010. Thus, as small as the Race to the Top 2011 grants were, the budgetary impacts of other federal education initiatives implemented during this time were almost certainly even smaller. In summary, it is highly unlikely that another education program caused the effects for the 2011 Race to the Top winners. No other major federal education program was implemented during the Race to the Top years, and even if there had been another program in parallel, this program would only explain my findings if it improved the 2011 winners more than the non-winners.

The second explanation is also somewhat unlikely. It is certainly believable that a state could benefit from thinking thoroughly about how to implement standards and accountability. And in developing a carefully laid out implementation plan, a state would also probably be more likely to implement these program components even after not receiving funding. However, with the meager extra funding that the 2011 winners received, this “willingness to implement” would need to almost singlehandedly account for the gains of one extra month of schooling per year that the 2011 winners saw in Math 4, Math 8, and Reading 4. The process of developing a careful implementation plan could certainly benefit the states, but an increase of an extra month of schooling per year would be massive for a program that assigned essentially no funding to the 2011 winners.

Explaining the gains made by the 2011 winners thus boils down to two unlikely accounts. But between the two accounts, the assumptions needed for the first seem much stronger than the assumptions needed for the second. As such, I am inclined to conclude that the application process itself led to the gains for the 2011 winners. In Appendix C I present a series of analyses that could in theory provide evidence for this “application process” effect, but these analyses have some methodological limitations that make them potentially flawed. As such, although my analyses from Appendix C could begin to suggest that the “application process” effect indeed is occurring, ultimately the only strong evidence in favor for the “application process” effect is the lack of alternative explanations for the gains seen by the 2011 winners.

To summarize, I find that the grant winners of both 2010 and 2011 saw improvements in their Math and Reading scores (with the exception of Reading 8 for the 2011 winners), but that the improvements were twice as large for the 2010 winners than for the 2011 winners. Furthermore, I find that only the 2010 winners saw improvements in their high school graduation rates. And since the 2011 winners had much smaller grants, their limited gains in test scores probably came thanks to developing a careful plan for implementing standards and accountability measures.

My findings on test scores align with the results from similar studies on accountability measures and standards. Previous studies have found that increased usage of standards, assessments, and accountability measures typically does translate to improvements in test scores (Lee, 2008; Hanushek and Raymond, 2005; Dee and Jacob, 2011), especially when test scores are directly linked to accountability measures (Deming et al., 2016). And since Race to the Top did succeed in increasing the use of standards and accountability measures in award-winning states (Dragoset et al., 2016; Howell, 2015), it seems natural that Race to the Top led to increases in student scores for states that implemented the program. My findings confirm this prediction, as I find that Race to the Top grant winners did manage to significantly increase students' test scores.

On the other hand, previous studies have found that standards, assessments, and accountability measures do not typically translate to improvements in graduation rates (Carnoy and Loeb, 2002; Deming et al., 2016). In this area my findings go somewhat against previous studies, as I find that the 2010 winners did improve their test scores by about 2% thanks to Race to the Top. However, since the 2011 winners did not improve their high school graduation rates despite improving their test scores, my findings still show that standards and accountability more easily improve test scores than graduation rates.

Limitations

As I mention in the Data section of this thesis, my findings on test scores are fundamentally limited by three shortcomings of the NAEP data. First, in the ideal world I would analyze all subjects, grade levels, and years tested by NAEP, but a significant portion of the NAEP data only has results at the national level. Historically Math and Reading have been the two main subjects analyzed by researchers, and they have often been used as a proxy for a students' overall academic achievement. Therefore, my usage of Math and Reading as proxies for overall student achievement is in line with the education world's standard research practices. Regardless, the lack of state-level data for other subjects certainly makes my evaluation of Race to the Top more limited. A second limitation related to the NAEP data is that only 4th and 8th graders were tested in Math and Reading, such that strictly speaking my findings only apply to these grade levels. It seems reasonable to assume that gains seen by 4th graders and 8th graders should be similar for other grades too, but only explicit analyses of student scores from these other grade levels would provide decisive evidence of Race to the Top's effects on these other grades. Third, the NAEP sample is demographically representative of the national population, but NAEP does not ensure representativeness at the state level. As such, if the state demographics drastically changed for most states in one of the groups (2010 winners, 2011 winners, and non-winners), then the demographic change could partially or fully account for the effects I find.

The high school graduation data also limited my thesis in important ways. First, my analysis of high school graduation rates depends on AFGR, a measure that is less consistent between states than ACGR. And although between-state differences would be controlled for in the state fixed effects of my regression analyses, the ultimate best would be for the measure to have both within-state and between-state consistency. Fortunately, the National Center of Education Statistics is now collecting ACGR data every year, so evaluations of future programs should not struggle with this consistency issue.

Beyond limitations of the data itself, my thesis is also fundamentally limited because I only examine Race to the Top's effects on student scores and high school graduation rates. Due to time and data constraints I concentrate exclusively on these two measures of student outcomes, but if the data available in a few years allows it, a post-program analysis of student outcomes should evaluate Race to the Top on other relevant student metrics. High school graduation rates are certainly a good estimator of future success because high school graduates earn much more than non-graduates (Greenstone et al., 2012), but other "success metrics" (college completion, marriage rates at age 30, salary at age 25, incarceration rates at age 25, etc.) would also be useful for evaluating Race to the Top's impact on students. Regardless, test scores and high school graduation rates are certainly a good starting point to begin estimating Race to the Top's direct impact on students. Standardized test data and graduation rates are often used by districts to measure school quality (Chicago Public Schools, 2020), so educators and policymakers certainly care about these two measures. As such, these two measures are useful metrics by which to evaluate Race to the Top's effects on students.

As mentioned before, standardized test data is often used to measure school or program quality. However, standardized test data inherently suffers from several limitations. First, standardized test data is only a proxy for student learning, and some researchers argue that test data should not be used to make high-stakes evaluations. Student scores sometimes fluctuate too much from year to year to be a reliable measure of class-wide learning, such that the test scores should not be used for determining program effectiveness (Koretz, 2017). Race to the Top is a program implemented throughout the nation and not just in a couple of classrooms, so for the purposes of my analyses these within-classroom fluctuations are not problematic. Regardless, it is still worth noting that one should be careful in using test data to establish effects on student learning.

The problem is not that standardized test scores are imperfect. All measures are lacking and subjective in some way. However, since test data certainly suffers from limitations just like all other measures, it should not be used as the only metric by which to evaluate a governmental program. In this thesis I expand beyond the test score data by also analyzing effects on high school graduation rates, but other measures are certainly also needed for a conclusive evaluation of Race to the Top. A comprehensive evaluation would need to look at a plethora of measures: standardizes test scores, attendance rates, core GPA, school climate, career outcomes, student/parent satisfaction with the educational system, among many others. The list of factors is certainly long and inevitably somewhat arbitrary, but looking at a range of outcomes (instead of just test scores and graduation data) is certainly necessary for a more holistic portrayal of the program's effects. Of these other measures deserving attention, teacher morale seems to be of particular importance. Several researchers have pointed out that the teacher evaluation system used under Race to the Top suffered from severe practical flaws, and both principals and teachers complained that under the new system they felt out of control of the metrics that later determined their contract renewal or termination (Winerip, 2011). It could plausibly be the case that Race to the Top produced changes that significantly lowered teachers' job satisfaction, and it is not immediately clear if small improvements in student scores and graduation rates are enough to justify such a negative effect on teachers.

In showing Race to the Top's effects on student scores and high school graduation rates I have contributed two vital criteria by which to evaluate the program's effectiveness, but these results are certainly just a fragment of the body of evidence needed to conclusively evaluate Race to the Top as a successful or failed program. And even in terms of determining the program's effects on test scores and high school graduation rates, my analyses are unfortunately not fully comprehensive. The implementation of standards and accountability sometimes leads to detrimental effects for the lowest performing students (Deming et al., 2016), so it could be the case that Race to the Top positively affected the high performers but negatively affected the low performers. If the group-aggregated gains were masking the program's detrimental effects on the lowest achieving students, then the gains seen by the rest of the students would be even larger than what I estimate. However, a disparity in group-specific effects would be problematic because it would mean the program increases inequity.

One way to test for the program's effects on the lowest achieving students is by looking at the effects on minority students, which are much more often in the lowest achieving quartile than their White peers. These race-specific analyses would certainly help shed light on Race to the Top's effects on students in different achievement levels. Unfortunately, the race-level data is too incomplete for test scores before 2010, such that the race-specific pre-treatment trends cannot be reliably estimated. And for graduation rates the race-level data simply does not exist before 2010. High school dropout rate data is available by race before 2010, but the dropout data cannot be readily combined with the graduation rate data. As such, I am ultimately unable to look at the Race the Top race-specific effects, and this limitation means that I cannot determine if Race to the Top was beneficial for students of all races and achievement levels. Future research should conduct analyses based on race or achievement level, either with newly released data from

NCES or with the data already available today. Currently the test score data by state and race is much more incomplete than the data just by state, but perhaps some analyses can still be run while controlling for the missing data. This endeavor seems like a good starting place for expanding the findings from this thesis.

Policy Recommendations

Race to the Top succeeded in increasing the use of standards and accountability measures (Howell and Magazinnik, 2017), and in this thesis I show that the program also helped improve student outcomes (test scores and graduation rates). Some researchers expressed concern about Race to the Top from the onset, arguing that by 2009 most states did not have the infrastructure (standardized tests, school accountability measures, teacher evaluation systems, etc.) necessary to implement the demands of the program (Haertel, 2009). And while the program does seem to have struggled with several implementation challenges (Winerip, 2011), the evidence from this thesis suggests that the program did succeed in creating some improvements within the educational system. Based on the findings from this thesis and previous studies, I propose the following five policy recommendations:

1. **Allow teachers and principals to divert from official recommendations when necessary.** The medical field began pursuing standardization far earlier than the educational field, and in its quest for standardization the medical field has left doctors with little capacity to divert from official recommendations (Timmermans and Berg, 2010). The overly prescriptive imposition of clinical guidelines has made physicians unable to use these guidelines in the most effective manner, as they are forced to follow the guidelines even in cases where the patients' individual characteristics make other treatment options more appropriate (Timmermans and Berg, 2010; Greenhalgh et al., 2014; Cook, 2015). Another challenge seen by the medical field during standardization is that the official guidelines are developed by outside actors with little on-the-ground experience (Greenhalgh et al., 2014; Howell, 2015, Bothwell et al., 2016,). Judging by the challenges faced during Race to the Top's implementation, it seems that the educational field is also struggling with these two challenges inherent to standardization. Obama stated that the

process of creating the teacher evaluation system should have explicit input from the teachers to avoid teachers being unfairly judged (the Obama White House, 2009), but Race to the Top simply did not engage teachers in the policy creation process (Winerip, 2011). Teachers could have helped anticipate the practical difficulties associated with implementing specific guidelines, such as the lack of test scores in certain subjects that fundamentally compromised school's capacity to properly implement value-added teacher evaluation systems (Haertel, 2009; Winerip, 2011). The negative effects of not engaging teachers in the program creation process could have been prevented by letting teachers and principals divert from the official guidelines when necessary (such as when test scores were unavailable for a certain subject). However, the overly prescriptive implementation of Race to the Top made diversion from official recommendations impossible. Thus, just as with the medical field, standardization in education has been an overly prescriptive process that outsiders impose on ground-level workers. To reverse this trend, future programs should ensure to engage teachers and principals during the policy creation process so that they can help predict the practical challenges of implementing the policy. And furthermore, since preparations can never fully envision all the difficulties that inevitably arise during a policy's implementation, teachers and principals should ultimately maintain the capacity to divert from official program guidelines when they deem it necessary. The program guidelines should act as recommendations of best practice, not as absolute requirements to be implemented even in cases where common sense clearly dictates otherwise.

2. **Combine goals for increased standardization with goals for a range of student outcomes.** Standardization and accountability can be crucial mechanisms by which to improve test scores (Hanushek and Raymond, 2005; Deming et al., 2016; Lee, 2008), but they need to be combined with other measures of student achievement (graduation rates, salary at age 25, etc.) if they are to

truly reflect improvements in the educational system. In and of itself, standardization of the educational field is a beneficial process because it increases transparency. And, ideally, standardization also leads to widespread implementation of the most impactful and cost-effective initiatives. However, the ultimate goal of education is to boost students' skills, experience, and knowledge (Lovenheim and Turner, 2017). A well-run public education system is a tool by which to ensure that students indeed gain skills and knowledge, but the system itself is not the end goal. Therefore, standardization is a process by which to directly improve systems and, hopefully, indirectly improve student outcomes. As such, education programs should have clear goals on both of these metrics and should develop detailed plans for evaluating the program's impacts on these two dimensions. Race to the Top only contained goals for system-level improvements, and in doing so the program was conceptually detached from education's ultimate goal of increasing students' skills and knowledge. In the future, even standardization-focused programs should have explicit goals for student outcomes, as student outcomes are certainly a necessary dimension by which to judge a program's impact on the education world as a whole.

3. **Maintain consistent measures of test scores and graduation rates.**

In conducting analyses on test scores and graduation rates, one fundamental challenge is that the measures often changed multiple times during the 1990-2015 period. For example, the AFGR data is available from 1990 to 2013, but after 2013 only ACGR is available. Due to this data limitation I combine these two measures into one, but it would be ideal to have continuous data on AFGR so that analyses can be conducted with that measure alone (even if AFGR is objectively a less desirable measure of graduation rates). And similarly, the test score data for most subjects is too incomplete to allow for thorough analyses. The Science test used by NAEP changed once in the 2000-2010 period, so the pre-Race to the Top trends cannot be reliably determined. And tests like Vocabulary and Writing were discontinued after only a couple years, so they also do not allow for longitudinal analyses. As such, NCES should ensure that the graduation and test data is more consistent throughout the years. Of course, the justification for changing the measurement mechanisms is that the new mechanisms are a more accurate representation of the construct measured, but as much as possible NCES should try to maintain measurements consistent from year to year. Keeping consistent measures will allow program evaluations to make less assumptions about how comparable the old and new measures are, which in turn allows for more airtight analyses.

4. **Implement more application-based programs.**

In this thesis I find initial evidence that Race to the Top's application process contributed to the gains later seen by the grant winners. If this evidence is indeed correct, then future education programs should consider implementing similar schemes for application-based fund allocations. The application process is a particularly cost-effective way to incentivize states into developing carefully planned implementation plans, and in this regard the application process itself can be a fundamental piece of creating improvements in the educational system. Thus, the education field should consider implementing more application-based programs in the future. And, importantly, evaluations of these programs can hopefully provide more evidence on the effects of including an application process as part of a program's rollout.

5. **When evaluating Race to the Top, conduct separate analyses for 2010 and 2011 winners.**

My analyses show that the states who won Race to the Top grants in 2010 received much larger grants than the states who won in 2011. As such, any study that investigates Race to the Top's impacts should look at the effects individually for these two groups. In this thesis I find evidence that these two groups in fact differed by how much they benefitted from the program, so future analyses on Race to the Top should certainly conduct separate analyses for the 2010 and 2011 winners. And more generally, evaluations of educational programs should acknowledge that the programs are not always implemented equally across states. If we look all program recipients as a single group (instead of creating subgroups based on implementation capacity or grant size), we lose the opportunity to determine the specific conditions most conducive to program effectiveness. And perhaps more importantly, null results found by previous researchers might instead be a significant effect for the 2010 winners that gets masked by the lack of results likely seen by the 2011 winners. Only looking at the 2010 and 2011 winners separately can allow us to confidently conclude program-wide null effects. As such, future evaluations of Race to the Top should certainly look at the program's effects on the 2010 and 2011 winners separately.

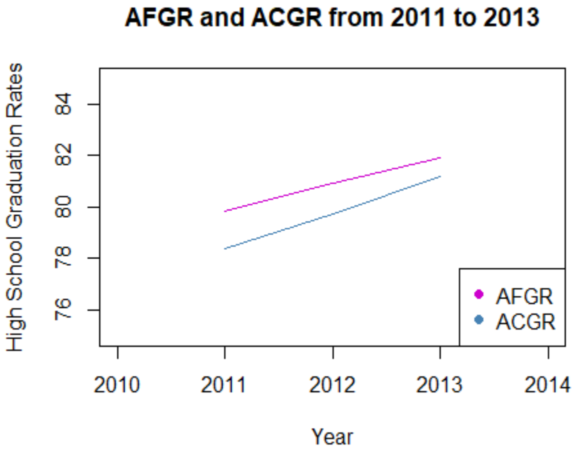
Conclusion

In this thesis I provide evidence that Race to the Top led to some improvements in student test scores and graduation rates. From the onset the program was aimed at increasing the usage of standards and accountability measures in the public education system, but based on my analyses the program also indirectly helped improve student outcomes. My findings therefore suggest that Race to the Top was indeed a successful policy, and although student outcomes are certainly not the only measure by which to evaluate the effectiveness of Race to the Top, the program's effects on test scores and graduation rates are certainly a useful place to start. Research has previously shown that the program did succeed in its explicit goal of increasing states' usage of standards and accountability measures (Howell and Magazinnik, 2017), and my analyses show that Race to the Top also had direct positive effects on student outcomes. Since my analyses contain a range of limitations I cannot decisively conclude that Race to the Top comprehensively improved student learning and rates of completing high school, but the evidence I provide certainly points in this direction. Thus, to the extent of the available data and methods, it seems that Race to the Top did lead to improved student outcomes. I can only speculate on the reasons for this improvement. Perhaps standards and accountability measures truly lead to improvements in instructional quality, which then translates to improved student learning. Standardization advocates would certainly argue in favor of this perspective, but future research should attempt to clarify if Race to the Top indeed led to improvements in teacher instructional quality. Regardless, with the research findings currently available we cannot make a decisive evaluation of Race to the Top as effective or ineffective, but the results from this thesis bring us one step closer to being able to do so.

Appendix A: Combining AFGR and ACGR into a Single Measure

Before combining AFGR and ACGR into a single measure, it is necessary to ensure that the two measures behave very similarly. In this Appendix I analyze AFGR and ACGR during the years 2011, 2012, and 2013. Since both measures allegedly measure the same construct (high school graduation rates), they should not show conflicting patterns during these years.

As an initial step for comparing AFGR and ACGR, it is useful to look at whether both measures follow roughly the same trajectory from 2011 to 2013. If both measures increase, decrease, or stay stagnant, then one would more readily accept that they indeed measure graduation rates in roughly similar ways. If, however, one of the measures showed significant decreases from 2011 to 2013 whereas the other measures showed significant increases, then one would suspect that the two measures are not in fact constructed in similar ways. Luckily, as the following graph shows, the two measures do indeed show the same trend from 2011 to 2013.

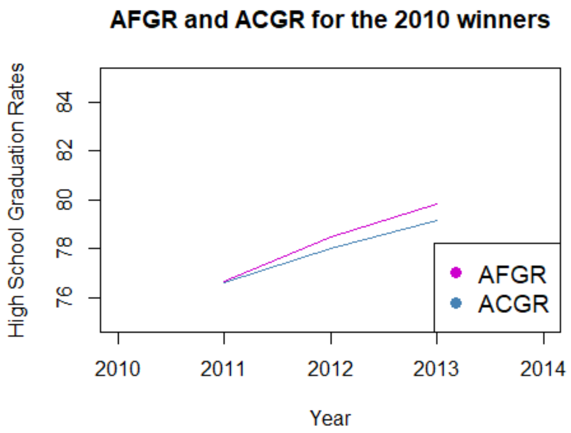


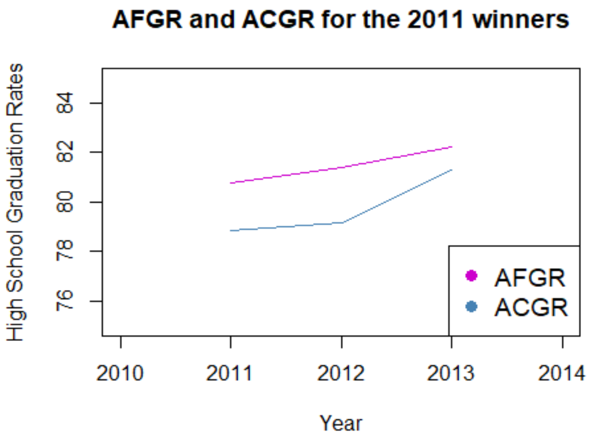
The two lines in the graph show the average AFGR/ACGR measure across all states during 2011, 2012, and 2013. And as can be seen in this graph, both AFGR and ACGR are showing increases of about 1% from 2011 to 2012 and from 2012 to 2013. Thus, the two measures are behaving similarly despite small differences in how the National Center of Education Statistics calculates them.

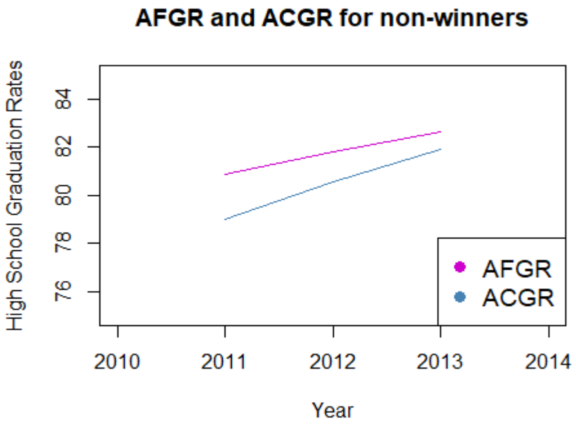
However, for the purposes of my analyses it is not enough for the two measures to behave similarly in the aggregate. Perhaps the two measures show increases in aggregate but show different patterns once subdivided into the three subgroups (2010 grant-winning states, 2011 grant-winning states, and non-winning states). My unit fixed effects regression analyses rely on the assumption that any changes that occurred to one group also occurred to the other groups, so if ACGR/AFGR behaves differently between groups, then my change of measure from AFGR to ACGR in 2014 could be creating false results. In other words, if the two measures do not work the same way for each of my subgroups, then any effect I find in my regression analyses could be at least partially occurring due to the change from AFGR to ACGR, rather than due to the Race to the Top grants themselves. As such, it is necessary to ensure that AFGR and ACGR also behave similarly for all three subgroups (2010 winners, 2011 winners, and non-winners) from 2011 to 2013.

Another important aspect to check is whether the difference between AFGR and ACGR remains consistent between groups. If one of the groups had significant differences between their AFGR and ACGR scores whereas for the other groups AFGR and ACGR are essentially the same, then my switch from AFGR to ACGR in 2014 and 2015 could be creating an artificial picture of decrease for the group with ACGR scores lower than their AFGR scores. And since the two measures are allegedly measuring the same construct, it is important that changing from one to the other does not unproportionally affect one of the groups. Thus, it is crucial to check that the difference between the AFGR and ACGR scores is about the same for all groups from 2011 to 2013.

The following graphs shed light on these questions about subgroup effects:


AFGR and ACGR for the 2010 winners

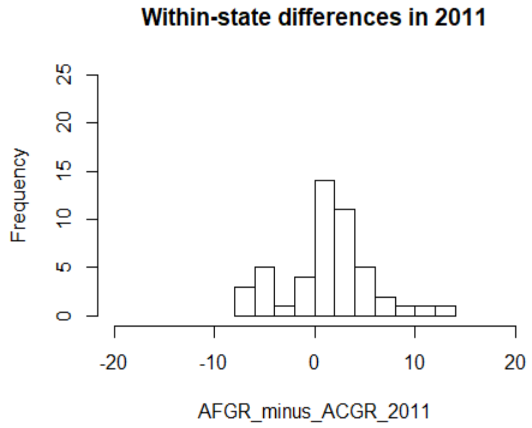

AFGR and ACGR for the 2011 winners

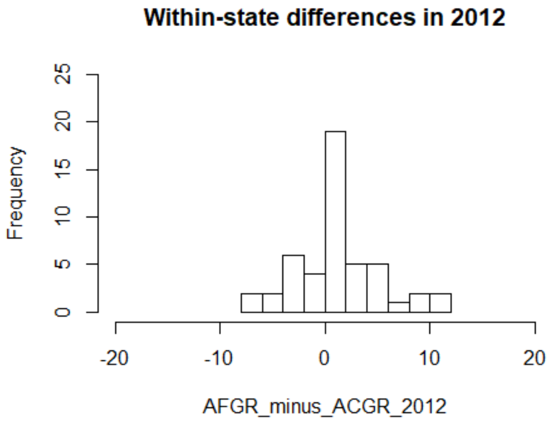

AFGR and ACGR for non-winners

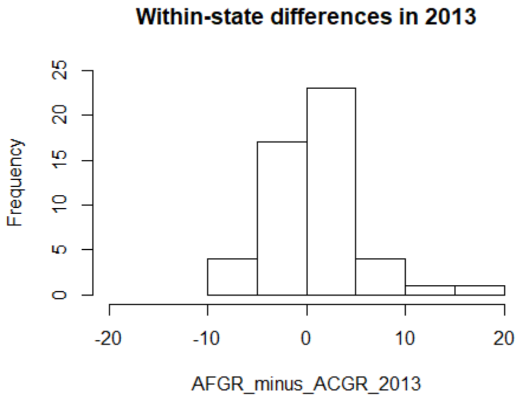
As can be seen in the graphs above, for all three groups (2010 winners, 2011 winners, and non-winners) both AFGR and ACGR increase from 2011 to 2013. And although the rates of increases slightly differ from subgroup to subgroup, for all groups both AFGR and ACGR increase by about 1% from year to year. As such, it seems that no subgroup effects exist for the rates of improvement captured by AFGR or ACGR.

One concerning aspect of the above graphs is that for 2010 winners the difference between AFGR and ACGR is essentially 0 during 2011. As the years go by the difference between the two measures slightly increases, such that by 2013 the 2010 winners also have AFGR rates that are slightly higher than the ACGR rates. However, whereas the 2011 winners and the non-winners both have differences between AFGR and ACGR that stay consistent throughout the years, for the 2010 winners the gap between the two measures seems to slightly change from year to year. The issue with this trend is that perhaps by 2014 and 2015 the 2010 winners would actually continue to see a widening gap between their AFGR and ACGR scores. If so, any alleged effects seen in 2014 and 2015 could in fact just be a reflection of the widening between-measures gap that occurs for the 2010 winners but not for the 2011 winners or the non-winners. Ultimately I believe that this trend does not fundamentally compromise my regression analyses because for 2010 winners the between-measures gap only widens by a small amount from year to year, but these subgroup patterns between AFGR and ACGR certainly work against my argument that the two measures can be readily combined.

One other important thing to check is whether the differences between AFGR and ACGR scores are small for each state. As I've showed up to this point, the measures behave very similarly to each other in the aggregate, and AFGR is in the aggregate only about 1 or 2 percentage points above ACGR (with this difference also remaining roughly consistent from subgroup to subgroup). However, this aggregated effect could be occurring if some states have much larger ACGR than AFGR while other states have much larger AFGR than ACGR. Based on the aggregated effect it seems that AFGR and ACGR are quite similar measures, but the measures could plausibly be working very differently from state to state despite looking similar in the aggregate. As such, it is useful to quickly check the spread of the within-state differences between AFGR and ACGR. The following graphs examine this spread:

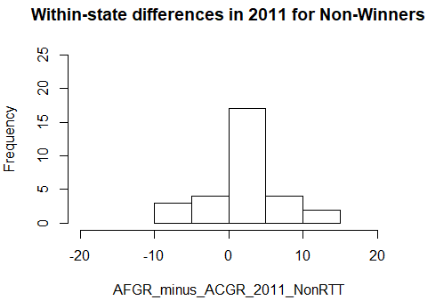


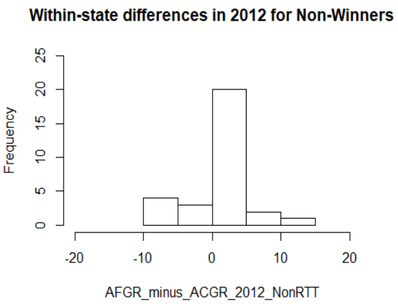


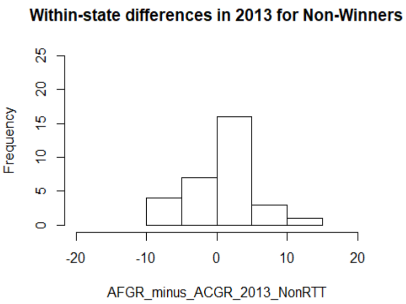


As these graphs show, for each year the within-state differences between AFGR and ACGR follow a roughly normal spread with mean close to 0. In other words, for most states the difference between their AFGR and ACGR scores are only a couple percentage points. Some outliers do exist where the AFGR score is significantly higher or lower than the ACGR score, but since the spread is roughly normal these discrepancies are not too worrisome. These holistic results are a sign that the two measures are roughly equivalent most of the time, but again it is necessary to ensure that this overall pattern is not hiding any subgroup effects (such as, for example, the 2011 winners consistently being the ones who have their AFGR scores much lower than their ACGR scores). The following graphs are constructed in the same way as the above graphs, but now looking only at the states from one subgroup at a time.

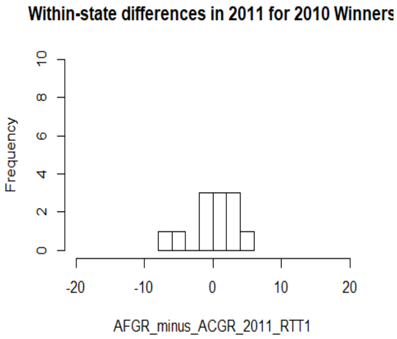
For states that did not win a Race to the Top grant:

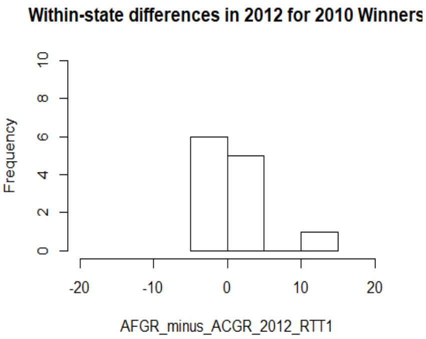

Within-state differences in 2011 for Non-Winners

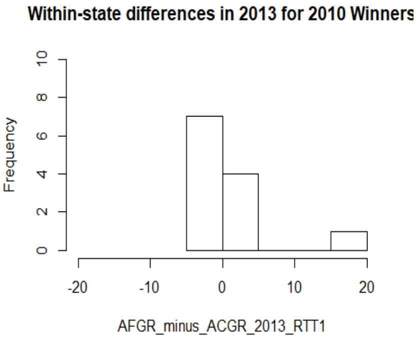

Within-state differences in 2012 for Non-Winners


Within-state differences in 2013 for Non-Winners

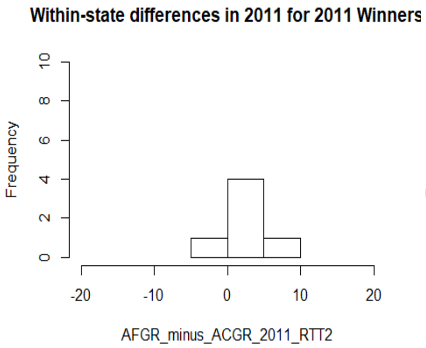
For states that won a Race to the Top grant in 2010:

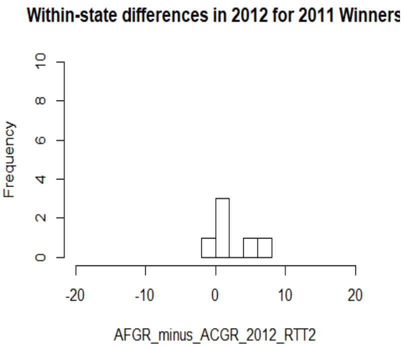

Within-state differences in 2011 for 2010 Winners

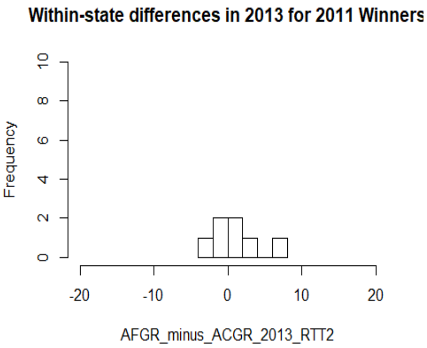

Within-state differences in 2012 for 2010 Winners


Within-state differences in 2013 for 2010 Winners

And for states that won a Race to the Top grant in 2011:


Within-state differences in 2011 for 2011 Winners


Within-state differences in 2012 for 2011 Winners


Within-state differences in 2013 for 2011 Winners

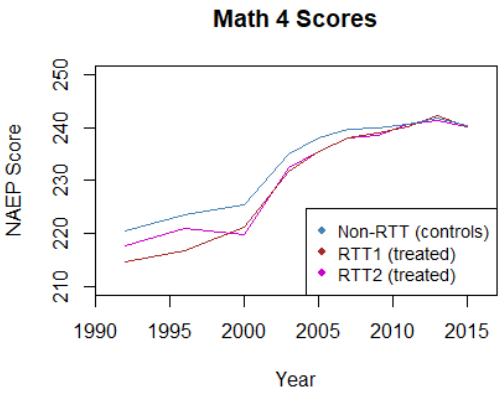
Because the sample size for some of the groups is so small (only about 10 states in the 2010 winners, for example), it is important not to overanalyze any one specific graph from the previous page. However, on the aggregate, these nine graphs show that most states within each subgroup had AFGR and ACGR scores similar to each other. This similarity between AFGR and ACGR is evidenced by the fact that for all graphs most of the states are at or close to 0, which is the score that occurs when AFGR and ACGR have exactly the same value. Importantly, no subgroup is showing a highly skewed spread. A skewed spread would be problematic because the measures are allegedly measuring the same construct, so for the most part they should have similar values for the same state within any specific year. Thus, it is reassuring that even within subgroups most states have AFGR and ACGR scores roughly 5 points from each other, and it is also reassuring that in aggregate the two measures are only a couple percentage points from each other (in average, AFGR is only about 1 to 2 percentage points above ACGR for any of the years examined).

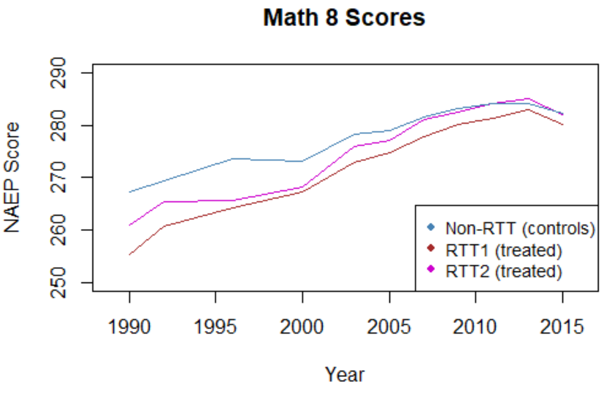
In conclusion, thorough analyses of AFGR and ACGR show that the measures can be reasonably combined as a single measure for the purposes of my unit fixed effects regression analyses. And importantly, I also ran my regressions with the AFGR data alone to ensure that any of my alleged results were also occurring when I maintained the same measure. With the AFGR scores alone I could only run the regression analyses until 2013, but even within this more limited timeframe the analyses showed that Race to the Top led to an increase in students' graduation rates. Thus, although the official findings reported in this paper are the analyses I conducted with the measure combining AFGR and ACGR together, the pattern of results would have stayed the same if I had chosen to conduct the analyses with the AFGR data alone.

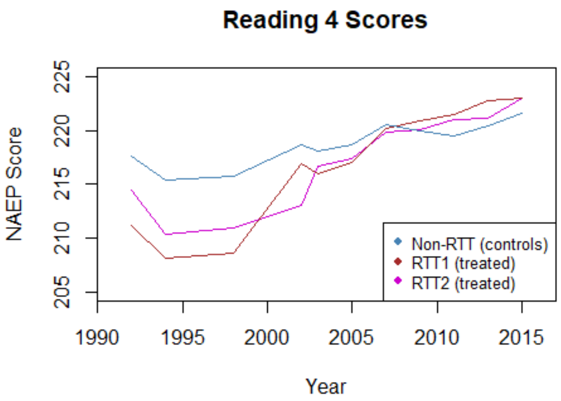
Appendix B: Testing the Linearity and Parallel Trends Assumptions

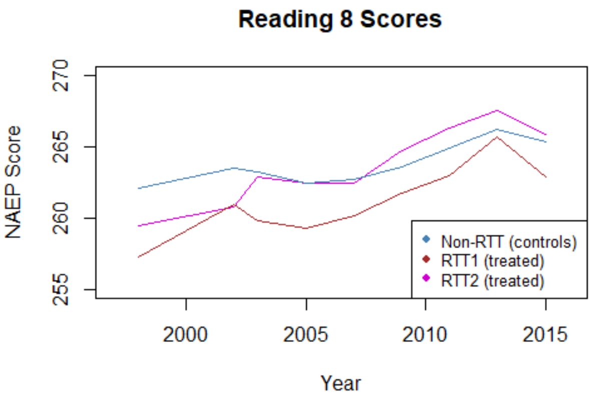
My unit fixed effects regression analyses rely on two fundamental assumptions. First, my model assumes that test scores and graduation rates were following a linear trajectory prior to Race to the Top. If the linearity assumption is violated, then my model severely misrepresents the values expected after treatment because the predictions would be based on linear trends rather than other trajectories (exponential, logarithmic, etc.). And second, my model assumes that all three subgroups (2010 grant-winning states, 2011 grant-winning states, and non-winning states) were following the same pre-treatment trends. The correctness of this assumption, known as parallel trends, is fundamental for ensuring the validity of the unit fixed effects regression analyses. When the parallel trends assumption is violated, the model cannot provide a reliable estimate of what future values should be after treatment. And in this way, a violation of the parallel trends assumption can lead to a fundamentally biased assessment of Race to the Top's effects. Thus, both the linearity and parallel trends assumptions must be fulfilled for my analyses to provide an accurate representation of Race to the Top's impact on student scores. In this appendix I evaluate if these two assumptions hold for the test scores and graduation rates data.

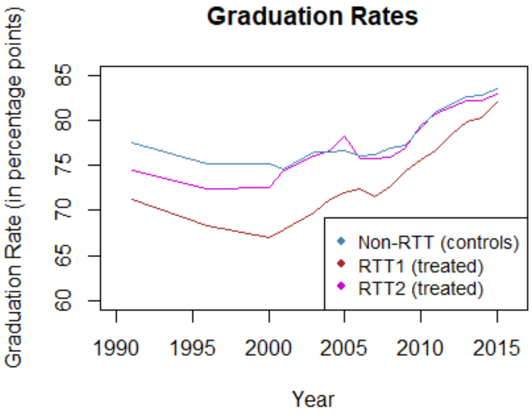
The following graphs show the trajectories of all three groups (2010 winners, 2011 winners, and non-winners) both before and after Race to the Top:


Math 4 Scores


Math 8 Scores







In all five graphs the lines represent the average performance of states within that subgroup. The brown line corresponds to the 2010 winners, the magenta line corresponds to the 2011 winners, and the blue line corresponds to the non-winners. As can be seen, the linearity assumption is well satisfied. In all five graphs the lines have a slight positive slope, and the increases seem to be well represented as following a linear trajectory. In other words, from year to year the test scores and graduation rates in each group seem to be increasing by about a constant amount. Perhaps the one exception to the linearity assumption is the slight dip seen by graduation rates from 1990 to 2000, as the trend during these years does not align with the increases seen from 2000 to 2015. However, since the graduation rates had been increasing from 2000 to 2010, and given that this year range is closer to Race to the Top than the 1990-2000 range, it still seems justified to say that the graduation rates had been linearly increasing before Race to the Top. Thus, as these graphs show, the linearity assumption is well satisfied for both the Math/Reading test scores and the AFGR/ACGR graduation rates.

Unfortunately, the case is less strong for the parallel trends assumption. For all five graphs (Math 4, Math 8, Reading 4, Reading 8, and high school graduation rates) the non-winning states started higher than the other two groups. In and of itself this difference between groups is not problematic, as the parallel trends assumption only requires that the increase rates are consistent between groups even if they start at different points. More problematic is the fact that, even before Race to the Top, the 2010 and 2011 winners seem to have been shortening the gap between them and the non-winning states. As such, it could plausibly be the case that my results are at least partially created by the factors that were shortening the gap even before Race to the Top, as these factors would presumably continue to shorten the gap during Race to the Top too. The problem is somewhat large for test scores, with the gap shortening by about five points every 5 years. Thus, the increases I calculate for the Race to the Top grant winners could plausibly be occurring due to this “shortening of the gap” rather than thanks to Race to the Top itself. For graduation rates the situation is less worrying, as the 2011 winners and non-winners had already converged by 2000 and continued to follow the same trajectory for the remainder of the years. And although the 2010 winners did have consistently lower graduation rates than the other two groups throughout this period, the shortening of the gap from year to year is quite small, such that even for this group the parallel trends assumption is still at least somewhat fulfilled.

One way to further check the parallel trends assumption would be to conduct placebo tests. In placebo tests, the same regression analyses are run for years in which no program was occurring. For example, a placebo test could call 2005 the treatment year and compare the before-2005 scores against the after-2005 scores. Since 2005 is a “placebo year” one would not expect any significant results to occur, as there’s no program during that specific year that would create gains. However, if the analyses show results even for this placebo year, then one would suspect a violation of the parallel trends assumption. Conducting these analyses could provide more evidence that the parallel trends assumption holds or does not hold for my regression analyses. As it stands I cannot ultimately conclude that the parallel trends assumption is fully satisfied for either the NAEP test scores or the high school graduation rates, and future research should conduct placebo tests or other similar analyses to ensure that the findings in this thesis are not being driven just by differences in pre-RTT trajectories followed by the 2010 winners, the 2011 winners, and the non-winners.

Appendix C: Potential Evidence for the "Application Process Effect"

In my analyses I find that the 2011 winners experienced some gains thanks to Race to the Top, even though the 2011 grants were small enough to be essentially inconsequential. More specifically, I find that the Math 4, Math 8, and Reading 4 test scores of the 2011 winners increased by about two to three points thanks to Race to the Top, even though the grants themselves only represented a 0.05% increase in the state education budgets. These set of results are certainly puzzling, and in this Appendix I present several paths of analysis that I pursue to try to create more evidence for my "application process effect" explanation. Ultimately, however, the analyses I attempt are inconclusive and cannot be used to prove the "application process effect".

If the "application process effect" indeed occurred, then the grant money should not be a perfect predictor of the gains seen by the winning states because some of these gains would have occurred through the "application process effect" instead. Therefore, if one finds that states with larger grants consistently saw larger gains than the states that received smaller grants, then it would seem that grant size drove the gains, and the case for the "application process effect" would be weakened. To shed light on this issue I conduct all my unit fixed effects regression analyses using a continuous treatment variable (Impact_Budget) instead of the binary interaction variable ReceiveRTT. I calculate the Impact_Budget variable through the following formula:

Impact_Budget = (Race to the Top Grant / 4) / 2010 state ed budget

For a full explanation about the components in this formula, I refer the reader to the Findings section of this thesis.

Using Impact_Budget instead of the ReceiveRTT interaction variable, I find that for every 1% education budget increase thanks to the RTT grants:

- **Math 4th grade** scores increased by 3.5 points for the 2010 winners (p<0.0005)
- **Math 8th grade** scores increased by 3.8 points for the 2010 winners (p<0.0005)
- **Reading 4th grade** scores increased by 4.2 point increase for the 2010 winners (p<0.0001)
- **Reading 8th grade** scores increased by 1.4 points for the 2010 winners (p<0.007)
- **High School Graduation Rates** did not increase for the 2010 winners (p-value not significant)

However, upon calculating these effects, it immediately becomes clear that the gains are about the same as when I use the regular binary ReceiveRTT as treatment variable. In other words, both the continuous Impact_Budget and the binary ReceiveRTT are showing effects of very similar size. Upon realizing this, I calculate the correlation between the two variables and find that they are in fact highly correlated with each other (r = 0.73). And because these two variables are correlated so highly, there's nothing to be gained from using the Impact_Budget variable instead of ReceiveRTT. As such, despite being continuous instead of binary, Impact_Budget lacks the variability necessary to disentangle itself from the general effects of receiving a Race to the Top grant.

Another potential path is to run the regression analyses with both Impact_Budget and ReceiveRTT at the same time. In this way, one can estimate the effects of larger grants irrespective of having received a Race to the Top grant in the first place. In other words, any gains that occurred thanks to winning a Race to the Top grant would now not be present for the Impact_Budget variable, and Impact_Budget would be a reflection of the grant size effect alone.

Unfortunately, because Impact_Budget and ReceiveRTT are so correlated with each other, they eat up each other's effect size when they are both included in the same regression. Specifically, the following are the results from my unit fixed effects regression including both measures instead of just one or the other:

- **Math 4th grade**: ReceiveRTT did not lead to a statistically significant score increase, Impact_Budget led to a 2.9 points increase for every 1% increase in state budget that occurred thanks to the Race to the Top grants ($p<0.007$)
- **Math 8th grade:** neither ReceiveRTT nor Impact_Budget led to a statistically significant score increase
- **Reading 4th grade:** ReceiveRTT led to a 3.7 point increase in scores ($p<0.0008$), Impact_Budget did not lead to a statistically significant score increase
- **Reading 8th grade** ReceiveRTT led to a 2.2 point increase in scores ($p<0.03$), Impact_Budget did not lead to a statistically significant score increase
- **High School Graduation Rates:** neither ReceiveRTT nor Impact_Budget led to a statistically significant score increase

From the above results, two points are worth noting. First, in most cases the variables are no longer acting as statistically significant predictors of gains, even though each was indeed significant when the other variable was not a part of the regression. And second, even in the cases where the variables do stay significant, the effect sizes are smaller than the effect sizes of the regressions with just one variable or another. This pattern of results suggests that the two measures are decreasing each other's predictive power, which occurs when measures are too closely related (Jones, 2016). For collinear measures the regression simply does not have a clear choice of which measure to use to account for effects, and thus the effects themselves are split between the two measures. Furthermore, the standard errors associated with ReceiveRTT and Impact_Budget double when the measures are both included into the same regression. This pattern also suggests high collinearity between the two measures, as standard errors typically increase when collinear measures are used in the same regression (Jones, 2016). Thus, ultimately ReceiveRTT and Impact_Budget are too collinear to allow for analyses evaluating the effect of grant size.

I experience a similarly fruitless endeavor when trying to look at the effects of having a higher application score. If the "application process effect" indeed occurred for the 2010 and 2011 winners, then the states with higher application scores should have seen larger gains in test scores and graduation rates than the states with lower application scores. The rationale is that a higher application score should in theory reflect a better plan for implementing the Race to the Top program components, and if the better implementation plans indeed led to some of the gains seen by the states regardless of grant size, then states with high application scores should see the largest improvements. Unfortunately, just like for Impact_Budget, using the application score as a continuous treatment variable is again futile because the measure is too highly correlated to the ReceiveRTT interaction variable (r = 0.78). Thus, just like Impact_Budget, the application score variable lacks the variability necessary to disentangle itself from the general effects of receiving a Race to the Top grant. Thus, due to lack of variation in both Impact_Budget and the application score variable, I am ultimately unable to conduct analyses that convincingly shed evidence on the existence of the "application score effect". Future studies evaluating application-based programs should attempt to test whether the application process itself leads to later gains, as the evidence I provide in this thesis is not compelling enough to confidently establish the existence of the "application process effect".

References

Alber, R. (2015). *Defining "Best Practice" in Education*. George Lucas Educational Foundation. Retrieved on April 1, 2020 from https://www.edutopia.org/blog/defining-best-practice-teaching-rebecca-alber

Baker, B. D., Oluwole, J., & Green, P. (2013). *The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era*. Education Evaluation and Policy Analysis Archives, 21, 1-71.

Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). *Assessing the gold standard—lessons from the history of RCTs*. N Engl J Med, 374(22), 2175-2181.

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

Chicago Public Schools (2020). *2019-2020 School Quality Rating Report*. Accessed January 7th, 2020.

Common Core State Standards Initiative (2019). *Development Process*. Retrieved on November 9, 2019 from http://www.corestandards.org/about-the-standards/development-process/

Cook, B. G. (2015). *How should evidence-based practices be determined*. Enduring issues in special education: Personal perspectives, 266-284.

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

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). *School accountability, postsecondary attainment, and earnings*. Review of Economics and Statistics, 98(5), 848-862.

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., ... & Upton, R. (2016). *Race to the Top: Implementation and Relationship to Student Outcomes*. NCEE 2017-4001. National Center for Education Evaluation and Regional Assistance.

Gabbay, J., & Le May, A. (2010). *Practice-based evidence for healthcare: clinical mindlines*. Routledge.

Gallivan, S. (2019*). The Impact of Integrating Student Learning Objectives in Missouri Teacher Evaluations on Student Academic Achievement.* Doctoral dissertation, Southwest Baptist University.

Greenhalgh, T., Howick, J., & Maskrey, N. (2014). *Evidence based medicine: a movement in crisis?*. Bmj, *348*, g3725.

Greenstone, M., Harris, M., Li, K., Looney, A., & Patashnik, J. (2012*). A dozen economic facts about K-12 education*. The Hamilton Project.

Haertel, E. H. (2009). *Comments on the Department of Education's Proposal on the Race to the Top Fund. Board on Testing and Assessment*. Washington, DC: National Academy of Sciences. Retrieved on April 18, 2020 from https://www.nap.edu/read/12780/chapter/1#10

Hanushek, E. A., & Raymond, M. E. (2005). *Does school accountability lead to improved student performance?*. Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management, 24(2), 297-327.

Harris, D. N., Liu, L., Barrett, N., & Li, R. (2020). *Is the Rise of High School Graduation Rates Real? High-Stakes School Accountability and Strategic Behavior*. Retrieved on April 18, 2020 from https://www.brookings.edu/wp-content/uploads/2020/02/Is-the-Rise-in-High-School-Graduation-Rates-Real-FINAL.pdf

Hill, H. (2009). *Evaluating value-added models: A validity argument approach*. Journal of Policy Analysis and Management, 28, 700–709.

Howell, W. (2015). *Results of President Obama's Race to the Top*. Education Next. Retrieved on April 18, 2020 from https://www.educationnext.org/results-president-obama-race-to-the-top-reform/

Howell, W. G., & Magazinnik, A. (2017). *Presidential prescriptions for State policy: Obama's race to the top initiative*. Journal of Policy Analysis and Management, 36(3), 502-531.

Imai, K., & Kim, I. S. (2019). *When should we use unit fixed effects regression models for causal inference with longitudinal data?*. American Journal of Political Science, 63(2), 467-490.

Ing, M. (2010). *Using informal classroom observations to improve instruction*. Journal of Educational Administration.

Jacob, R. (2016). *Using Aggregate Administrative Data in Social Policy Research*. OPRE government report. URL. Retrieved on April 18, 2020 from https://www.acf.hhs.gov/sites/default/files/opre/opre_brief_draft_dec2016_finaldraftjacob_clean_508.pdf

Jacob, R., Somers, M. A., Zhu, P., & Bloom, H. (2016). *The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions.* Evaluation review, *40*(3), 167-198.

Jones, K. (2016). ResearchGate. *What's the link between Multicollinearity and standard error (S.E)? Should it occur in any regression model?* Retrieved on April 13, 2020 from https://www.researchgate.net/post/What_the_link_between_Multicollinearity_and_standard_error_SE_should_it_occur_in_any_regression_model

Klein, A. (2015). *No Child Left Behind: An Overview*. Education Week. Retrieved on November, 2020 from https://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.

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

Lohman, J. S. (2010). *Comparing no child left behind and race to the top*. Connecticut General Assembly, Office of Legislative Research.

Lovenheim, M., & Turner, S. E. (2017). *Economics of education.* Macmillan Higher Education.

Miller, T. & Hanna, R. (2014). *Four Years Later, Are Race to the Top States on Track?*. Center for American Progress. Retrieved on April 1, 2020 from https://www.americanprogress.org/issues/education-k-12/reports/2014/03/24/86197/four-years-later-are-race-to-the-top-states-on-track/

Moran, D. J., & Malott, R. W. (2004). *Evidence-based educational methods*. Elsevier.

NAEP (2018). *About NAEP*. National Center for Education Statistics. URL. Retrieved on November 16, 2019 from https://nces.ed.gov/nationsreportcard/about/.

NCES (2020). *National Center of Education Statistics: Home Page*. National Center of Education Statistics. Retrieved on April 2, 2020 from https://nces.ed.gov/

NCES (2020). *Trends in High School Dropout and Completion Rates in the United States.* National Center of Education Statistics. Retrieved on April 2, 2020 from https://nces.ed.gov/programs/dropout/ind_04.asp and https://nces.ed.gov/programs/dropout/ind_05.asp

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). *Learning to learn from benchmark assessment data: How teachers analyze results*. Peabody Journal of Education, 85(2), 226-245.

Ripley, A. (2013). *The smartest kids in the world: And how they got that way*. Simon and Schuster.

Robinson, W. S. (1950). *Ecological correlations and the behavior of individuals.* American Sociological Review, 15(3), 351–357.

Sass, E. (2020). *American Educational History: A Hypertext Timeline.* Retrieved on April 10, 2020 from http://www.eds-resources.com/educationhistorytimeline.html.

Somers, M. A., Zhu, P., Jacob, R., & Bloom, H. (2013). *The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation*. MDRC.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). *Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment.* American Journal of Evaluation, *35*(3), 311-327.

Tatter, G. (2015). *As Tennessee finishes its Race to the Top, teachers caught in the middle of competing changes*. Chalkbeat. Retrieved on March, 2020 from https://www.chalkbeat.org/posts/tn/2015/12/15/as-tennessee-finishes-its-race-to-the-top-teachers-caught-in-the-middle-of-competing-changes/

The Obama White House (2009). *President Obama on Race to the Top*. Retrieved on April 18, 2020 from https://www.youtube.com/watch?v=VNbDv0zPBV4

Timmermans, S., & Berg, M. (2010). *The gold standard: The challenge of evidence-based medicine and standardization in health care*. Temple University Press.

Timmermans, S., & Mauck, A. (2005). *The promises and pitfalls of evidence-based medicine*. Health Affairs, *24*(1), 18-28.

U.S. Department of Education (2016). *Race to the Top Fund.* Retrieved on April 1, 2020 from https://www2.ed.gov/programs/racetothetop/index.html

Winerip, M. (2011). *In Tennessee, Following the Rules for Evaluations Off a Cliff.* New York Times. Retrieved on March, 2020 from https://www.nytimes.com/2011/11/07/education/tennessees-rules-on-teacher-evaluations-bring-frustration.html