4-2019

Identifying Exceptionally Performing Third Grades in Indiana: Who Is to Blame?

Emily Troyer 19
DePauw University

Identifying Exceptionally Performing Third Grades in Indiana: Who Is to Blame?

By: Emily Troyer

DePauw University Honor Scholar Program, Class of 2019

Sponsor: Dr. Humberto Barreto, Q. G. Noblitt Professor of Economics and Management
Committee Members: Dr. Jamie Stockton, Associate Professor and Chair of Education Studies; Dr. Steven Bogaerts, Associate Professor of Computer Science

Abstract: Federal and state governments want to know that their constituents' tax dollars are being spent well, and that public education is serving the People well. To asses a school's efficacy, they calculate an accountability score. Although state-level scoring systems vary, neither the federal law nor a single state's law require that demographic characteristics of a school's student population be held constant when it is being evaluated. This complicates the assessment of a school's adequacy because factors outside of a school's control influence student performance; when a school is evaluated without taking this into account, it is being unfairly credited with its students' successes and failures. I use OLS regression, holding some population characteristics constant, to create predicted weighted average ISTEP scores for third grades in Indiana to compare to their actual scores. The deviation between actual and predicted scores more accurately reflects how well or poorly the school is educating its students. I propose an alternate, more statistically rigorous grading system which identifies exceptionally performing schools. This approach is by no means foolproof, but it certainly gives a better foundation for assessing a school's influence on student outcomes than a grade based upon a school's average standardized test score. This current approach controls nothing about student population, while my method gets us closer to a fair evaluation of a school's performance.

TABLE OF CONTENTS

***

ACKNOWLEDGMENTS

First and foremost, this thesis would not have been possible without Professor Barreto. He has provided honest, insightful feedback on this thesis from the very beginning (after so kindly agreeing to sponsor me when I insisted I needed him specifically). Perhaps more importantly, though, the ideas he espouses in class, the seeds he plants in his attentive students' minds, his "never settle" mentality, and his passion for educating have deeply inspired me in this thesis and beyond. Applying economic thought to non-economic problems is one of the most important things we can do with economics, and without B I would not have thought to apply economics to this education dilemma in which Indiana and the rest of America find themselves. Thank you for sitting through my brainstorms (read: word vomits), for answering my numerous questions, for sending me thought-provoking articles and podcasts, and for pushing me when I needed that extra shove, both on this thesis and throughout my time at DePauw. We've come a long way since I cried after taking your intermediate microeconomics midterm, and I like where we've ended up.

To the rest of my committee, Professor Stockton and Professor Bogaerts: thank you for agreeing to sacrifice your time and energy without really knowing me. I appreciate your votes of confidence and the technical insights with which you have provided me throughout this process. (And a thank you to Professor Sayili and Professor Raghav who answered my many questions on STATA usage.)

To Allison Roehling, my trusted advisor: thank you for giving me feedback on a number of thesis topic ideas that eventually led me to this one, and for first inspiring my love of economics in intermediate macroeconomics. My last semester at DePauw has felt strange without you.

To Toph: thank you for being my sounding board for all of my sometimes nonsensical ideas, for helping me filter and distill them, and for listening to many iterations of what eventually became this thesis. You have supported me every step of the way and I cannot thank you enough for that. Also, thank you for reminding me to take a breather every now and again and enjoy the present.

To the Honors Scholar department: thank you for providing this opportunity for us to explore our intellectual passions. The entirety of the program, from freshman through senior year, has been one of my favorite parts of college. Thank you to each and every faculty member from whom I have had the pleasure of being challenged, especially Jeff Kenney, Beth Benedix, and Dr. Foss. You all, in different ways, have shaped who I am today and how I think. You have all modeled what it looks like to question truths, to search for meaning beyond what is immediately before our eyes, and to contemplate alternative options to the current norms.

To Cathedral High School's international baccalaureate program: thank you for giving me a space to be curious, and for showing me the joys of pursuing intellectual interests outside of class. Mrs. Bradshaw, thank you for developing my passions for writing; both you and Buffy have set brilliant example for what a lifelong learner looks like. I hope one day I can impact people to the extent you both do.

And finally, thank you to my parents. Your sacrifices, the care with which you have raised me, and the tough love you have shown me over the years have allowed me to take full advantage of my educational opportunities. It is because of the multitude of positive educational experiences with which you have provided me that I am driven to more deeply contemplate how we can work to improve the experiences of other students in Indiana.

I. INTRODUCTION

Part A: Opening Remarks

"We bad, huh," an Arlington High School senior asked WFYI correspondent Eric Weddle in 2015 after he noticed Weddle had been to their school with recording equipment multiple times (Weddle, 2019). Principal Law had stopped this particular student that day, but not other students, and the kid was displeased. Since the takeover, enrollment at Arlington High School had dropped further, funding had all but disappeared, and the company hired by the State Board of Education to captain the takeover broke its contract. Principal Law had too many fires to put out to give each one the same amount of water. Despite being taken over four years prior, the causes of Arlington High School's takeover had not improved according to Principal Law. Absenteeism, pot, and vandalism were only a few of the hourly issues faced by Principal Law. Teachers left mid-term—some even left before the semester began—and seventh graders were reading at the level of a nine-year-old. One Arlington student, several friends and relatives of students, and the boyfriend of a student were all killed during the 2015-2016 school year. Still, the state gave Arlington High School an ultimatum: get its graduation rate to 60% by the end of the 2015-2016 school year, or face a shut-down—but how can a school be judged for things it itself cannot fix?

The current structure of accountability grading and punishment draws an unfair conclusion on how a school is performing. An elementary school's accountability score is judged based upon 1) the ISTEP scores of students in that year, and 2) the change in scores from year to year. Both measures are reliant upon ISTEP scores, which, as is explained below, depend upon more than just the school itself. Why punish a school for factors that are outside its control? This main issue with Indiana's current method of holding elementary school's accountable can be vastly improved by

simply holding factors outside of a school's control constant, and grading elementary schools based on an appropriately adjusted ISTEP score.

Abernathy (2008, p. 25) sums up this issue:

> Parents send their children to school for the better part of their waking lives but rarely have much of an idea about what has happened at the critical interface between their minds and the variety of experiences they encounter and in which they engage… The child's teacher—if he is attentive and responsive—probably has a better idea about how well the student is doing than anyone else… Chances are that the principal does not know more than the teacher… The principal's administrative superiors probably know even less… they assign standardized tests to the students… Any accountability program… must accurately measure what a given school or teacher is adding to the knowledge and skills of a given student, above and beyond [external factors].

The accountability grading system is intended to evaluate how a school performs. A school's performance is contingent upon the students, teachers, and administrators at that school. Teacher and administrator performance are within the school's control, but student performance is a byproduct of the teaching they receive along with a myriad of other factors including some that cannot even be quantified. The goal of this thesis is to quantify and hold constant factors that affect a student's performance to produce a score that accounts for factors outside of the school's control so that comparison between third grade performances in Indiana are fair and funding can be more properly allocated.

***

Part B of this section describes the background for accountability scoring, and what the stakes of these scores are. Section II part A delves into literature investigating my conjecture that a not insignificant part of student performance is dictated by factors unrelated to school, the importance of third grade to a child's long-term educational success, and general theories regarding testing and how well it evaluates students. In part B, the methodologies used by the other forty-nine United States to calculate accountability scores are reviewed in order to look for models

which hold constant external factors when assigning accountability scores. The final part of this section, part C, describes theory and calculation behind the testing scores that are a crucial part of Indiana's current accountability grading system. Part D of Section II explains the dependent variable of this thesis—the ISTEP scale scores and my manipulation of these scores. Part E explains how the Indiana Department of Education (IDOE) calculates its accountability scores for elementary schools, and by extension how they would calculate a score for the third grade. Section III describes the source of the data, how I constructed the data set I used, and the methods of my thesis. Section IV includes the various regression models I used, my alternate grading system, the results, and the implications and conclusions of these models. Finally, Section V will tie up the findings of this analysis, discuss the implications of this research in public policy, and suggest future avenues of study.

Part B: Accountability Scoring and Its Stakes

> "Teachers that remain continue to build relationships and focus on academics, but some students have complicated, even dangerous, lives." –Eric Weddle, WFYI's "The Takeover"

i. A History of National Education Mandates and Reforms

President Johnson signed the Elementary and Secondary Education Act (ESEA) into law in 1965 to show the American people that the federal government was committed to ""quality and equality" in educating… young people" (Brenchley, 2015). Society realized the powers of stronger educational systems from President Truman's 1950s campaign against communism and to increase national security; the public sphere turned to education as a means of restructuring the socioeconomic disparities in the US between racial groups. The act became a key component to

Johnson's "War on Poverty," despite its marginal returns (Jeffrey, 1978). There are two types of grants available to school districts per Title I of the ESEA: a "basic" grant and a "special incentive" grant. Basic grants can be given to any district which has a certain number of low-income families. Special incentive grants can be given to any district which qualifies for basic grants and also is implementing programs and procedures to specifically improve the achievement of the children of low-income families; the district must be making attempts to meet student needs (ESEA, 1965; "The ABCs of ESEA…", 2019). Both grants are available to school districts across the nation up to certain values (ESEA, 1965).

This act required state schools to show their Department of Educations that measurable objectives (which came to be known as annual measurable objectives, or AMO) in school performance were being achieved. Part B (ii) of this section expands on AMO. Based on these state evaluations, the federal government doled out funds to schools and districts with high levels of low-income students to provide aid in the following areas: "professional development, instructional materials, resources to support educational programs, and the promotion of parental involvement" (Paul, 2016). Not only did the ESEA mandate stringent academic guidelines and measurement of improvement, but also that the outcomes needed consequences ("The Big Idea of School Accountability," 2015). Report cards detailing the fulfillment of these guidelines are required to be created and submitted to the federal government under section 1111 of the ESEA (ESEA, 1965).

President George W. Bush renewed and reformed the ESEA in early 2002 through what eventually became a highly controversial bill, despite its initial bipartisan support. It is called the No Child Left Behind (NCLB) act. Like the ESEA, part of this legislation's aim was to increase the efficiency of schools. Furthermore, it sought to mitigate persistent achievement gaps between

low-income students and high-income students (Abernathy, 2008, pp. vii-viii). This was the first piece of national legislation that demanded improvement in teaching and outcomes for historically low-performing groups such as minorities, students from a low socioeconomic status, and students with special needs (Chatterji, 2006, p. 490). The NCLB also required annual testing in all states and required that states use these scores to evaluate the performance of each school ("The Big Idea of School Accountability," 2015). Instead of testing once in elementary school and once in middle school in only some states, all states' grades three though eight were required to take standardized tests each year (Gewertz, 2015). Finally, the NCLB mandated transparency in accountability scoring so that parents, educators, and taxpayers alike could understand how their schools are performing. Despite lofty goals, the framework for accomplishing these unprecedented feats was ambiguous (Chatterji, 2006).

In 2012, the Obama administration implemented the Every Student Succeeds Act (ESSA), another reauthorization of ESEA ("Every Student Succeeds Act (ESSA)," n.d.). It relaxed the imposition of federal standards (Brenchley, 2015) for states who could show "rigorous and comprehensive…plans to close achievement gaps, increase equity, and improve the quality of instruction" ("Indiana's 2015 Flexibility Waiver…", 2015). The ESSA was signed into law in December of 2015 and began its implementation in 2016. State education agencies are now free to "request flexibility on behalf of itself, its local educational agencies (LEAs), and its schools, in order to better focus on improving student learning and increasing the quality of instruction" ("Indiana's 2015 Flexibility Waiver…", 2015).

A major goal of the ESSA was to reintroduce flexibility into the management of public schools so that schools could be run according to their individual needs while continuing to be accountable for student outcomes ("Every Student Succeeds Act…", 2018). Because of backlash

from the NCLB, particularly the feeling that it implemented standards too rigorous and consequences too severe for many schools, this malleability was a key point in the ESSA. The ESSA maintains the annual testing requirement for grades three through eight, but loosens the parameters on what this test must look like, allowing for "portfolios… projects, and performance tasks" to act as assessments (Gewertz, 2015). States must continue to publish report cards as previously mandated in the ESEA ("Overview of the State Accountability Report Card," 2018). They are not required to compare the results in the report card to and therefore base grades on AMO as they were in the NCB era; the standard is now "state designed long-term goals" (A. Whelan, personal communication, December 18, 2015).

One of the most notable aspects of the ESSA, however, is that it abolishes provisions which impose federal consequences for poorly performing schools and districts (Gewertz, 2015). This strikes out a key component of both the ESEA and the NCLB. There are, however, still state imposed consequences which will be discussed in part (iii) of this section. The elimination of federal consequences of poor performance was controversial, with critics fearing that progress made under the NCLB will diminish and supporters rejoicing in the restoration of power to the states. States now have "greater discretion to implement… evidence-based improvement strategies and interventions" ("Overview of the State Accountability Report Card," 2018), these supporters say.

ii. Impact of Annual Measurable Objectives (AMO) on the Accountability Grade

AMO are submitted by the state to the federal government in order for them to approve the benchmark each state is setting for its grading system. More specifically, they are the annually-determined "target for the percentage of students whose test scores must be proficient or above in English/language arts and mathematics" (Ritz, 2017, p.4). In order for a school to receive an "A"

grade from the federal government, it must have reduced achievement gaps consistent with the state's achievement goals; these goals must be "applied consistently throughout the state for all public schools, districts, and subgroups of students" (https://eddataexpress.ed.gov/definitions.cfm). For elementary schools in Indiana, this meant that the performance indicator must improve. The IDOE defines student subgroups as "economically disadvantaged students, major racial & ethnic groups, students with disabilities, [and] students with limited English proficiency;" these subgroups of the student population must independently meet the AMO for the school to be eligible for an "A" grade (McCormick, n.d.). The advent of ESSA, as highlighted in part B (i), has done away with AMO and therefore the explicit, tangible objectives schools must meet in order to receive certain accountability grades from the federal government. However, Indiana will continue to have its own grading system it uses to assign state accountability scores (M. Paino, personal communication, January 28, 2019). I do not focus on the federal grades and how it impacts Indiana state policy. The system I am focusing on is the state's system of accountability, as this impacts whether the state takes over schools and how the state allocates its funding.

The AMO subgroups of the student population are subgroups that I am hoping to account for mathematically; my regression analysis and prediction of an average ISTEP score for each third grade, holding these subgroups constant, should give a measure of how the school is doing. The current accountability scoring system crosses its wires in trying to improve the education of disadvantaged student groups, while still punishing the school for the effects of these student groups on their own test scores. The improvement of the performance of subgroups is vital. Due to the partially structural and systematic nature of the negative performance effects of these subgroups, though, schools themselves can have only a muted influence on subgroup

improvement. Schools are not being accurately compared if different schools have different numbers of disadvantaged students.

iii. The Stakes

There are real consequences for continuously poor performing schools. A school which received an F is immediately at risk for a downgrade in accreditation status. The repercussions are in title only—the state does nothing. However, if the F's persist until there have been at least four total, and a probationary accreditation is awarded, the school and its corporation have one year to raise their grade from an F. If this does not happen, then the Indiana State Board of Education (ISBE) hosts a "public hearing in the school corporation where the school is located to consider and hear testimony concerning… options for school improvement" (IC 20-31-9-4(b), 2018).

Matthew D. Harsanyi, Accreditation Specialist at the IDOE, says that this means that if a school does not improve its accountability score above an F after a year, it and its corporation is at risk of an IBSE takeover (M. Harsanyi, personal communication, January 31, 2019). The occupation is viewed as necessary because local authorities who were charged with running a school adequately are evidently incapable of doing so and should therefore be removed. However, residents of local school districts elect the local boards that run the schools. The people decide who runs their district. When the state seizes a school, the people are no longer represented in the operations of the school.

A little more than half of a school's funding comes from the state and federal government; the local school board of each district is responsible for raising funds for the remainder of their budget. This is usually done in the form of collecting property taxes on businesses and homes in the district. The fund collection, in the event of a takeover, can be delegated from the school board

to the city council for that district so as to maintain local involvement in budgetary operations ("How Are the Local, State, and Federal Governments Involved…", 2019).

Additionally, the IBSE could decide to merge the poor-performing school with another, or close the school altogether (IC 20-31-9-4(b)). Both of these options create a burden on a neighboring school and potentially increase the cost of transport for the district or for families themselves. If the option chosen by the ISBE for schools which have received four consecutive F's and have been unable to turn this designation around is not closure or merging, then it will be designated as a turnaround school (IC 20-31-9-4(c)). State, local, and federal tuition support funding will be pulled from the school corporation in which the turnaround school resides. In the event that a charter operator takes this school over instead of a state-appointed team, then the funding will be redirected from the corporation to the charter operator (M. Harsanyi, personal communication, January 31, 2019).

It is vital, then, that schools are awarded the most appropriate grade possible. The consequences of a misplaced low grade are dire. Before schools are punished for poor performance, everything possible must be done in order to ensure the school is receiving acclaim or blame for only that which it is responsible. It is my hope that this thesis allows us to come up with even a small improvement to accurately predict the grade the school should receive because of only its own effects.

II. LITERATURE REVIEW

> "The past quarter-century has seen no gains in overall student performance at 17. Gains observed at 14 dissipate by the time students reach the last year of high school… Whether and how these gaps can be narrowed is beyond the scope of our study. But it's clear that what America has been doing, at a cost of hundreds of billions of dollars, hasn't worked." –Eric A. Hanushek and Paul E. Peterson, *WSJ* March 17, 2019

Part A: Factors Influencing Student Performance and Standardized Testing

This section will review research on factors influencing student performance, the importance of third grade, and general theories regarding testing and its effectiveness in evaluating students. There exists a massive body of literature regarding influences on student academic performance. There are a multitude of factors, from the location of the school to the socioeconomic status of the student, that are outside of a school's influence but nevertheless impact student performance. The importance of achievement in third grade is justified in this section to reflect the purpose this paper's focus on analyzing third grade ISTEP scores in Indiana. A few of these drivers are expanded upon in this section to show the importance of attempting to control for some of these drivers when assigning accountability grades to schools in Indiana. The advantages and disadvantages of standardized testing are evaluated in order to understand the use of standardized test scores in ultimately assigning accountability grades. Finally, special emphasis will be placed on the notion of achievement gaps as I attempt to show why controlling for individual differences when grading schools statistically is more effective than current norms.

i. Achievement in Third Grade

Hernandez (2011, p. 5) found that of the children not graduating on time from high school, 16% of those are not reading at grade level by third grade. Furthermore, "one in six children who are not reading proficiently in third grade do not graduate from high school on time, a rate four times greater than that for proficient readers" (p. 3). These are alarming figures, and hopefully act as a wake-up call for those who do not view elementary school as a crucial time in a person's life. Third grade is also a decisive time in the foundation of a student's ability to learn skills which apply reading; it is in third grade when students begin to read in order to learn, and are no longer

simply learning to read, according to Hernandez (2011, p. 4). After third grade, he says, "interventions for struggling readers… are seldom as effective as those in the early years" (p. 4).

Frisby (2013) supports Hernandez (2011)'s premise; Frisby (2013) says grades one through three focus on learning to read, and from then onward students use reading to learn. Students who are not capable of reading on grade level at the time this transition happens—from third to fourth grade—will begin to fall behind. Additionally, Frisby (2013) cites third grade as the year that students will "begin to lose their enthusiasm for school," and failures during this time tend to stick with students into their academic careers (p. 216). Getting the proper allocation of resources from the state is essential for third graders. This requires an acutely accurate accountability grading system and is why I have chosen to hone in on third grade accountability scoring in Indiana schools.

ii. Factors Influencing Standardized Test Scores

In shifting gears from third grade to factors influencing school achievement, I turn to Sibley and Dearing (2014) study in which they compared the involvement in early education of immigrant parents to that of US-born parents, and investigated the achievement impacts of their levels of involvement on their children. Positive associations between educational involvement and student achievement were found by Sibley and Dearing (2014) for "US-born white, black, and Asian families" and especially for Latino immigrants.

They used Epstein (2011)'s organization of family educational involvement (FEI) subcomponents for their study. The components were broken down into six types by Epstein (2011): "(1) parenting and basic obligations such as ensuring the child's safety; (2) communication between home and school; (3) in-school activities such as volunteering in the classroom; (4) helping the child to learn at home through directly engaging in learning activities and a stimulating

environment; (5) decision making such as serving on a Parent-Teacher Association committee; and (6) collaboration with the community," along with parent expectations regarding academic achievement and rule implementation (Sibley and Dearing, 2014, p. 814). These domains have consistently shown to have great direct and indirect impact on achievement of children, Sibley and Dearing (2014) claims. Additionally, the domains of parental expectations and school involvement have shown the strongest associations with achievement. Educational involvement from parents has the strongest positive impact for students who are at a socioeconomic disadvantage, and may even "compensate for disadvantages in the home environment and social capital of families" (Sibley and Dearing, 2014). This involvement is much easier for well-off families (Chatterji, 2006).

When Sibley and Dearing (2014) analyzed the rates of educational involvement for both first and third grade immigrant and non-immigrant children. The area in which the most difference existed between immigrant parents, specifically Asian, Black, and Latino ones, and of US-born white parents was in their involvement related to school. This result could surely influence the achievement gap between student subgroups, and is not something schools have much control over. These findings support the more general literature they cited and which I summarized above.

Haskins and Jacobsen (2017) also holds that parental involvement in elementary school is important for many child outcomes. The authors' definition of parental involvement incorporates "any learning-related effort provided by a parent or caregiver to increase their children's educational outcomes" (p. 658). Many potential hurdles exist to parent involvement, including, as Haskins and Jacobsen (2017) cite, "parent and child traits, economic constraints, family circumstances, demographic characteristics, teacher characteristics, and school contexts" (p. 659).

However, they reiterate the meaningfulness, albeit partiality, of parental involvement as a predictor of academic success (p. 659).

This study goes a step further by specifically focusing on the negative impact that paternal incarceration has on children's educational outcomes. The authors found that a father's incarceration inhibits their involvement at home and in school, and also that the mother's involvement is lower in families which have an incarcerated father. Haskins and Jacobsen (2017) do, however, acknowledge significant issues of reporting bias, as involvement is difficult to define, and even more so to quantify (p. 661). Their research reinforces Sibley and Dearing (2014)'s and my own hypothesis that external factors outside of a school's control significantly impact student performance. If I had access to clean measures of both parental involvement and incarceration, I would include them in my study.

Another determinant of student performance is the location, in the context of urbanness versus ruralness, of the school and thus its students. Bæk (2016) maintains that rural education needs more attention, as local community influences social belonging (p. 441). She begins by noting that when independent variables like socioeconomic status and parental educational attainment are controlled for, location has an independent effect on a child's success in school (p. 437). The variability in the performance of rural students is a "recurring theme" in international studies (p. 435). She goes on to say that a place is more than just its physical location, but it aggregates a certain set of individual characteristics that play out differently in one space over another. Regional differences are important to consider because of the fact that different types of people are living in and drawn to live in those different settings (Bæk, 2016). Put differently, Bæk (2016) says that "regional variation is related to differences in the characteristics of those living

there" (p. 437). Furthermore, Bæk (2016) posits that a student's choices and motivations vary with the specific conditions and possible barriers specific to a locale.

The issue she sees, though, is the different opportunities and resources available to urban schools that are not present in rural schools. Bæk (2016) relays that urban schools are assumed to be the standard. The possibility to live out the "urban ethos" preferences that many youth possess, regardless of their own location, is less likely to come to fruition for rural youth. Young people can feel this narrow possibility, and this acknowledgement may in turn affect their "motivations to learn, their beliefs about their own abilities and their learning strategies" (Bæk, 2016, p. 440). In fact, the system of education itself makes it difficult to imagine staying in a rural setting, as the main focus of education is not on the needs for the occupations for rural labor markets. These findings by Bæk 2016 solidified the importance of taking into account important differences between rural and urban schools. Locale is a variable I wish I could have controlled for in my models, but the data provided by the IDOE was properly formatted for my usage and its categories were arbitrarily assigned and relatively meaningless.

Bæk (2016) also comments on gender performance in school. Girls tend to do better in academic settings and more often attend higher education than boys. Males are also more concerned with the present, while females are more likely delve into schoolwork and to look to the future. This is likely to the masculine dominance of immediate spaces, Bæk (2016) thinks. Females are less attached to a specific location. This is an interesting contrast to the typical political rhetoric, which insinuates the disadvantage females are at in higher education and in education in general. This can be applied to the discussion on women's underrepresentation in STEM fields, which Bæk (2016) did not evaluate.

There are a multitude of studies, in fact, which delve into gender differences in school. Much of the contention in gender differences in schooling and the reasons for studying them lie in what Bæk (2016) briefly points out, which is how women and men perform comparatively on math and English or reading sections of standardized tests. Fryer and Levitt (2010) empirically show that early in schooling, a gap emerges in math between males and females in favor of males. Spread throughout 1,000 schools, they use a sample of more than 20,000 students who are followed from kindergarten to fifth grade. Within these six years, boys perform significantly better than girls on an item response theory (IRT) test in mathematics. The gap is 0.2 standard deviations in favor of males by the end of fifth grade, compared to no mean gap upon beginning school. Conversely, Chatterji (2006) finds that males are 0.31 standard deviations below females in reading abilities at the end of first grade. He uses a cohort of 2,296 students from 184 schools across the U.S. who were tested before entering kindergarten and at conclusion of first grade. The students are members of the Early Childhood Longitudinal Study (ECLS).

Husain and Millimet (2009) address the mainstream media's claim that boys are losing ground to females in the first four years of school. They say that while this may be true for average achievement gaps, it is quite clear that "boys outperform girls in math across virtually the entire distribution by the end of third grade, and gain ground across the entire distribution over the first four years of school" (p. 39). This finding is only statistically significant across the distribution, though, for whites. Even in reading, where Husain and Milliment (2009) as well as previous literature show female dominance in reading before kindergarten and at the end of third grade, most boys do not lose ground to girls; boys at the lowest end of the reading-ability spectrum are the only ones who lose ground during that time period.

Gender gaps in math performance emerge prematurely, in pre-kindergarten to be precise, according to Cimpian, Lubienski, Timmer, Makowski, and Miller (2016). They find the gap first occurs for high achievers, and then proliferates through the distribution of abilities. Moreover, Cimpian, Lubienski, Timmer, Makowski, and Miller (2016) found that teachers perceive the abilities of males and females with similar achievement levels differently; male math proficiencies are rated higher than females with comparable achievements and learning behaviors. They also note that learning approaches between genders differ throughout the distribution, with girls' studious habits paying off the most at the lower end of the achievement spectrum. As gender is a contentious issue in education, I will be using it as an independent variable in my study.

Another variable affecting test score outcome is whether or not a student is learning English as a second language. As previously mentioned, third grade is when students are expected to use reading to learn other things. If a student is still learning to read and understand the language, it will be difficult for her to use the language to learn and to show what she has learned. Valdez-Pierce (2003) points out that often times, a traditional assessment's result for an ELL conflates that student's intelligence with her language ability. Thus, it is difficult to interpret an ELL's ability from a standardized test. The goal of the IDOE in targeting ELLs is to reduce their achievement gaps, implying that with better instruction an ELL can do better on the test. However, Valdez-Pierce (2003) finds that ELLs achieve these "standards-based learning goals" and at a slower pace (p. 11). While she admires the goal of the ESEA, she does not believe schools can be realistically expected to meet its goals; this is especially true when sampling and measurement error are considered (Valdez-Pierce, 2003). Nevertheless, Valdez-Pierce (2003) remarks on the importance of assessing ELLs through standardized tests and maintaining high standards of achievement for them.

Valdez-Pierce (2003) particularly notes that the ESEA achievement gap closure standards, which Indiana also possesses, force ELL students to "outperform general education students." This, too, is unrealistic. If schools with many ELLs are performing poorly and lacking funding, then they will likely not have the resources for effective bilingual programs which Valdez-Pierce (2003) suggests for ELL students who are not progressing. The issue of ELL achievement gap closure, then, becomes cyclical.

iii. Evaluating Standardized Test Scores

Indiana Statewide Testing for Education Progress (ISTEP) scores are the dependent variable of this study, and the foundation of the component that comprises the IDOE accountability grade itself. Because I do not intend to discard the base of the current grading system, standardized testing itself deserves some discussion. The way that testing outcomes vary for certain subgroups and the cause of this variation is directly tied to whether or not a school is responsible for the variation of subgroups' test scores. Caldas and Bankston (2005) report that black and Hispanic students historically perform worse on measures of academic achievement, namely standardized tests like the NAEP, SAT, and ACT, than white and Asian students. They cite many reasons for this, one major reason being a lack of social capital among disadvantaged children who are "disproportionately black and Hispanic in the United States" (p.195). Social capital is defined by Caldas and Bankston (2005) as "healthy networks of interactions between children, parents, neighborhoods, and schools" (p. 195). This is one reason I include race and Hispanic ethnicity in my model.

Hispanic and black children are also more likely to be food insecure and go to bed hungry; poor nutrition is strongly correlated with poor academic outcomes. They are also more likely to live in antiquated, dilapidated buildings, painted with lead-based paint which is linked to juvenile

learning problems and decreased IQ's (Caldas and Bankston, 2005). This can negatively affect one's cognitive ability reflected on standardized test scores. Black children are also watching higher amounts of television than their white and even Hispanic counterparts, which has a negative relationship with academic performance (Caldas and Bankston, 2005). Therefore, the racial, ethnic, and socioeconomic composition of schools should be held constant when assigning an accountability score to a school, as the demographic composition of the school is not something that can be, or should be, altered. Playing with incentivizes to change the demographics, when data so poignantly reminds us of gaps of achievement between minorities and majorities, reminisces of days pre-1964.

Jencks and Phillips (2006) ascribe the gaps in test scores between blacks and whites to other environmental factors. Blacks score lower than whites and Asians on tests that measure both material taught and not taught in schools, and the achievement gap is not larger when the tests "appear to measure familiarity with the content of "white" or "middle-class" culture than on tasks that appear to be more culturally neutral" (Jencks and Phillips, 2006, p. 83). Furthermore, it is difficult to measure, in advance, most of what affects performance in a job or in college. Cognitive skills are the most easily, most accurately measured traits that predict performance, and it is this predictive measure on which blacks are most disadvantaged. Jencks and Phillips (2006) point out that performance-based evaluation benefits blacks, whereas test-based selection, which emphasizes only certain cognitive skills, puts them at a disadvantage.

Jencks and Phillips (2006) conclude that standardized tests have "harmed blacks as a group," but that they are not flawed (p. 84). Differences in skill measured by standardized tests are real, and these skill differences have meaningful real-world application. The inability to measure other performance indicators, they say, is the biggest social problem regarding performance

predicting assessment as these other indicators are ones which blacks are far less disadvantaged on. Test-based skills do not fully predict the performance of an individual. This is a rather intuitive conclusion, and for that reason, perhaps an overhauling of the main component of Indiana's accountability grade through a shift from ISTEP scores is necessary. This is not what I attempt to accomplish, however, so understanding the drivers of test score differences and the efficacy of test scores serve as important intuition behind my models.

Further in their investigation, Jencks and Phillips (2006), account for over half of the test score gap between black and white five- and six-year-olds who took a standardized test through a set of "family environment indicators" (p. 104). They note that this was difficult because these characteristics could be a proxy for both genetic composition and environment. The indicators are as follows: "grandparents' educational attainment, mothers' household size, mothers' high school quality, mothers' perceived self-efficacy, children's birth weight (a proxy for prenatal environment), children's household size, and mothers' parenting preferences" (Jencks and Phillips, 2006, p. 104). They cite parental education as one of the best indicators of socioeconomic status disparities. This shows the importance of disadvantaged environments for minorities in test score outcomes, which schools have no control over.

Three more critical determinants of standardized test performance were delineated by Grodksy, Warren, and Felts (2008) in their sociological report. They are "learning, cognitive ability, and opportunity to learn (OTL)" (p. 386). Learning is a change process over time, the authors say, by which a person acquires information and skills. Cognitive ability, as previously discussed in this review, is much more challenging to define. As Frisby (2013) points out, there is the *g* (general mental ability) factor which mostly defines intelligence and is discussed later in this

section; Grodksy, Warren, and Felts (2008) describe that anywhere from two to sixty other "hierarchically nested cognitive abilities following the *g* factor" exist for humans (p. 388).

Debate ensues over the extent to which intelligence is unchanging or heritable; either way, the purpose of standardized tests—to test either achievement or ability—have become linked in American testing (Grodksy, Warren, and Felts, 2008). They define OTL more simply as "the resources available to students, most often in classroom settings, that facilitate their acquisition of knowledge or skill" (Grodksy, Warren, and Felts, 2008, p. 388). This is an important input to testing results because OTL's efficacy is contingent upon a student's ability to learn, which itself is contested. However, it is undeniably critical to distill and improve this input as much as possible; this is what schools can control, and this is the component of ISTEP scores I hope to reveal through the regression models.

Average standardized test scores, Grodksy, Warren, and Felts (2008) found, do vary along racial and ethnic boundaries and by socioeconomic status. They importantly note, though, that variation within groups is often larger than variation between groups. Despite this, it is the inter-group variation that is of most interest to researchers because it indicates "important differences in home and school resources as well as possible flaws in the tests themselves" (Grodksy, Warren, and Felts, 2008, p. 387).

Grodksy, Warren, and Felts (2008)'s research is supported by Chatterji (2006). He reports that these "environmental factors such as family and school supports" can minimize the probability of academic failure for students whose group identification, such as race, gender, or socioeconomic status, puts them at risk for failure. Chatterji (2006)'s conjecture extends to the intra-group variation in test scores pointed out by Grodksy, Warren, and Felts (2008), implying that it can be alleviated through supportive home and school environments. If home support data was easily

quantifiable, I would include it in my models to see what kind of influence they have on third grade ISTEP test scores in Indiana. Since it is not easily available, I will not be using it. I did, however, want to emphasize the importance of both inside and outside of school environmental factors on standardized testing.

iv. Achievement Gap

Intimately linked with standardized testing is the issue of the achievement gaps between subgroups of students themselves, and the government's policies directed at closing them. The policies and accountability scoring in general were discussed in Section I part B. Because the closure of these achievement gaps are a core goal of evaluating schools, it is important to analyze what aspects of these achievement gaps should be credited to a school's performance. Test score gaps between subgroups, which effectively become achievement gaps for the purposes of public rhetoric, are multifaceted.

The concept that a child's performance and a school's performance are based on more than what can be measured is not an original idea. Abernathy (2008) specifically addresses the shortcomings of legislation like the NCLB and the complexity of, and therefore potential problem in, measuring educational quality through his book *No Child Left Behind and the Public Schools*. In this book, he ascertains two foundational considerations regarding the evaluation of a school: "What determines good educational quality?" and "How, if at all, can we observe and identify it?" (p. 27). By answering these questions, or rather not being able to fully answer these questions, Abernathy proves the hazard of current accountability systems and legislation.

He posits that there is likely no system that perfectly measures a school or teacher's performance due to the uncertainty of circumstances that contribute to educational success. Furthermore, he asserts that convoluted aspects of how accountability systems and laws, like

NCLB, affect "the production of education" are too often overlooked in lawmakers' and researchers' quests for student improvement; since federal mandates and even state mandates are, at their deepest, concerned with schools, not the individual children or teachers or even classrooms (p. 27). Principals, the executers of schools, bear the burden of running a school compliant with federal legislation. They are key to a valuable educational experience, but "can only do so much, bound as they by a web of competing demands and situations" (Abernathy, 2008, p. 28). This idea forms the premise of my thesis.

Legislation like the No Child Left Behind Act (NCLB) does not acknowledge that the measures by which it assesses schools, and those by which many states assess schools (including Indiana), are factors that lie outside of a school's control (Abernathy, 2008). Furthermore, "omitting any attempt to fix or even acknowledge that resources and communities matter for academic achievement, NCLB assumes" that schools alone can rapidly and dramatically improve academic achievement for all students, disadvantaged and not (Abernathy, 2008, p. 29). Disparities between the quality of school matter more for disadvantaged students (Abernathy, 2008). Therefore, being able to draw out the real quality of the school by controlling for student performance influencers outside of the school's control will provide invaluable insight into the true differences in quality between third grade schools in Indiana. Then, the state government can more properly focus on closing true quality gaps in schools which would more effectively help disadvantaged students.

Finally, Abernathy (2008) discusses how public schools can be at a further disadvantage because of behavioral externalities in school systems. Disruptive behaviors in school are examples of "congestion effects—difficulties that arise when the inability to exclude people from receiving benefits of a public good [like education] results in" overuse of the service (Abernathy, 2008, p.

30). Abernathy (2008) cites the most obvious form of congestion effects in education as the disruptive behaviors of one student which decrease quality education for and teacher attention on the other students. He also points out that while private schools can combat congestion effects by awarding scholarships to exemplary students and threatening suspension or expulsion to disruptive ones, public schools lack these more forceful tools. Because standardized tests results "capture mostly students' experiences outside of the production process," an accountability system based upon these results should be weary of other ways to evaluate school effectiveness (Abernathy, 2008, p. 31). I mentioned previously that I plan to work within the current structure that Indiana uses to evaluate schools by using standardized test scores, but noting congestion effects is important when considering what is and is not inside a school's control but may affect student performance on these exams. If it were possible given my research constraints, I would have included congestion effects as a control variable in my models.

In a similar vein to Abernathy (2008), Barton and Coley (2009) concern themselves with the feasibility of closing an achievement gap. In his original 2003 study, Barton says that "gaps in school achievement, as measures, for example, in the eighth grade, have deep roots—deep in out of school experiences and deep in the structure of schools. Inequality is like an uninvited guest who comes early and stays late" (Barton and Coley, p. 46). This time, Barton and Coley (2009) performed a meta-analysis of research, and conglomerated sixteen factors which relate the broad life of a student and his or her cognitive development and achievement. They went on to analyze how these "correlates of achievement" vary between racial, income, and ethnicity groups (Barton and Coley, 2009, p. 3).

Their results reveal that the correlates do vary between groups, and that the differences do in fact mirror the achievement gaps seen in academic achievement. Barton and Coley (2009)

conclude that closing the achievement gap first requires the closure of gaps in the life experiences of these various subgroups—an important conclusion policymakers should internalize. The achievement gap may be narrowed or maintained, but so many contributors to student performance lie outside of a school's control that it would be illogical to credit the school with all of the success or failure of its students' performances.

General cognitive ability is a determinant of student success that Frisby (2013) asserts is the single most impactful factor in understanding a student's capacity to comprehend academic. This, however, has become a controversial idea among academics and politicians; Frisby (2013) goes so far as to invert a famous George Orwell quote by saying that "some ideas are so obvious that only an intellectual would deny them" (p. 201). This concept of general cognitive ability as the most vital way to assess how well a student learns in an academic setting is one of those ideas. Unfortunately, this concept strains "the opposing ideals of meritocracy versus equality" in society (Frisby, 2013, p. 201).

The *g* (general mental ability) factor is a persistent and strong source of error in any sort of broad cognitive test—from one third to one half of variance (Frisby, 2013). Frisby (2013) goes on to claim subgroup intelligence (IQ) testing and standardized testing do have substantially different averages; for IQ testing, he finds like the mean score for whites, American blacks, Hispanics, East Asians, and Native Americans are 100, 85, 90, 106, and 90 respectively. All subgroups achieve the full spectrum of IQ scores. Finally, as far as mental testing is concerned, Frisby (2013) ascertains that different means and standard deviations between subgroups are not adequately explained by biased or unfair tests. Therefore, standardized test results between student subgroups can in part be attributed to expected *g* factor, and raises a question relevant to this thesis: for how much of the *g* factor are schools responsible?

With this in mind, Frisby (2013) claims that "never before has the chasm between established scientific research and political wishful thinking been so wide as in the contemporary rhetoric on "closing the achievement gap"" (p. 212). There are several documented cases in which pouring money into a project to close this elusive achievement gap has produced no results, and in which the project itself had no empirical evidence to suggest that it would succeed in its endeavors. To suggest that one can just "close the achievement gap" through educational policy targeted at schools is to deny individual differences in cognitive ability; even if the entire population were the same race and ethnicity, individual differences would "remain and are an inescapable fact of life" (Frisby, 2013, p. 213). The degree to which these differences stem from environmental or biological determinants is a separate question, but the differences themselves exist (Frisby, 2013).

This isn't to say that interventions and altered instruction cannot help both students and teachers deal with differences in ability, but that when individual differences exist, measures of improvement will increase the performance of higher achievers at a steeper rate than lower achievers (Frisby, 2013, p. 213-214). The rhetoric, then, should be changed to "close the achievement gap within individuals," since many students do not work up to their potential in academic environments. Any and all students' performances can be improved with the "right opportunities, effort, and instruction" in conjunction with their respective abilities (Frisby, 2013, p. 215).

To understand Frisby (2013)'s arguments, look no further than the Chicago Public Schools (CPS). Hood (2011) writes that CPS are almost half black students, most of whom are from low-income households. The issue of closing the achievement gap is something important to this school system. Unfortunately, CPS has not been able to share in the national trends towards greater increases to black students' performances on the National Assessment of Educational Performance

(NAEP)[1] than white students' performances since the 1990s. Within the CPS school system, black students' performances are improving, but at a slower rate than white ones, Latino, and Asian students. Hood (2011) notes many confounding issues here, including extremely high suspension, expulsion, and disciplinary rates among black students as well as a 50% graduation rate, which has been increasing but not at the same pace as other racial or ethnic groups in the CPS system.

Hood (2011) also points out that Chicago, unlike New York which is also notorious for poor public school systems, has neighborhoods still largely segregated by race and economic status; the root of the achievement gap extends beyond the school and must be addressed there as well (Hood, 2011). This point is one of the most important points that my thesis attempts to address when grading schools, since almost a quarter of third grades in Indiana are a majority students of color.

As far as testing in the context of achievement gaps is concerned, though, Helms (1992) is concerned with the lack of formal models to account for cultural factors, particularly ones specific to respective racial and ethnic groups, present in cognitive ability tests (CATs). The researcher defines these CATs as "measures designed to assess intelligence, mental abilities, cognitive abilities, and scholastic aptitude" (p. 1). The premise of Helms (1992) is that environmental and biological factors, which are often used to explain away black and white gaps in CAT performances, are not a complete story. They fail to take into account the effects of exposure to racism which Helms (1992) found to be a statistically significant contributor to gaps between white and black students.

[1] The NAEP is a national standardized cognitive ability test which fourth-graders and eighth-graders take across the country. This data is often used to compare, on a national level, ability among students.

This is particularly interesting to my thesis because this could indicate issues with the ISTEP test itself, the test upon which accountability grades are based, and could explain why achievement gaps are slow to close. This again is something which schools can only govern in a limited hand. Unfortunately, I will not be able to control this as Helms (1992) did, but it is something of which to be aware. It should be emphasized that these previous summaries serve solely to show the problematic rhetoric and goal setting involved with accountability scoring, nationally and, by extension, in Indiana. My empirical analysis begins to reveal the true condition of students' opportunities to learn as demonstrated through their ISTEP scores. In an ideal world, I would standardize the populations between schools in order to control for these elusive, arguably immeasurable qualities like intelligence.

Part B: A Fifty-State Comparison of Accountability Scores

At minimum, the ESEA and its reauthorizations require all states to report the condition of their schools as discussed in Section I part B (i). All states, including Indiana, also use the accountability scores to discern the level of state involvement necessary for specific schools and districts, which is discussed in Section I part B (iii). Before proceeding with my relatively simple regression model for improving Indiana's third grade accountability scores, I wanted to find out if any other states use regression in accountability scoring or otherwise attempt to hold constant other factors that affect accountability scores. The short answer is yes, one single state does in fact use regression to hold constant factors that contribute to its accountability score. However, no state *creates* its accountability score by holding constant important demographic factors outside of a school's control like my models. ("50 States Comparison," 2018).

Evidence for this somewhat surprising result comes from guidelines for each state's accountability scoring. The Education Commission of the States, the actor for James Bryant

Conan's original concept of a cooperative on states' education policy, compiled a table summarizing the accountability systems for almost every state's grading policies ("History," 2018). Using this and the links to each state's accountability site, I examined and summarized the grading systems for each individual state. This full analysis can be found in Appendix A. The conclusions of this investigation were curious: not a single state used regression to account for demographic variants in its calculation of schools' effectiveness.

New Mexico is the lone state that uses regression to calculate any of the components of its EMS accountability scores. For the Current Standing component of the score, the regression holds constant school size, student mobility, and previous scores in order to get a predicted subcomponent result for the overall score. The two error terms in the model—one for random school effect and one for random student effect—comprise the residual, which is the difference between the school's actual and predicted score. This residual is then standardized, transformed into a zero to one probability via a cumulative normal distribution function, and then multiplied by the maximum possible point values of this component. Similar processes are used for the other two components. Generally, the goal of New Mexico's models is to measure how far above or below the school's actual scores are from their predicted scores, and use a normalized version of this to contribute to the total points available for each component (Ruszkowski, 2017).

While the factors chosen by New Mexico to include in the accountability scores are indeed held constant, they are largely unrelated to student-specific characteristics. Some other factors which New Mexico considers are previous Math and ELA scores, whether the student was a full-year student, whether the student took an alternative assessment, the number of students who took assessments, and the percentage of students who are not full year academic students; notably, none of these variables have anything to do with a student's inherited circumstances. Many states

attempt to account for social or demographic factors by tacking on bonus points of sorts to the final score, but none truly account for these factors in the way that regression does. The approach used in this thesis is applied in a novel way: to get an accurate read on how well a school is actually running by trying to hold constant the influence of things outside of a school's control.

Part C: ISTEP Scale Scores

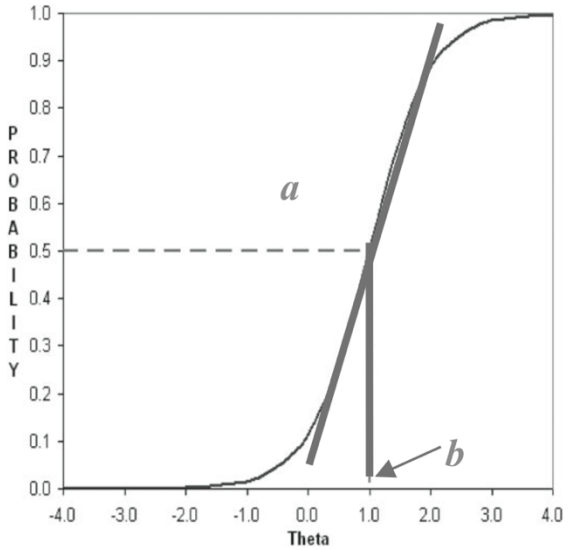
Indiana Statewide Testing for Educational Progress (ISTEP) exam results reflect "to what extent an individual has mastered the Indiana Academic Standards in the English/Language Arts, Mathematics, Science, and Social Studies content areas" (Indiana Department of Education "ISTEP+", 2017, p. 1). This means that the ISTEP test is meant to gauge a student's acquired skills, not his intelligence. Below is a brief explanation as to how ISTEP[2] is scored and what this scoring style means. Understanding this is important because third grade ISTEP scores form the dependent variable in the regressions in this thesis. Additionally, ISTEP scores are the sole determinant of the Growth and Performance indicators which comprise elementary and middle school accountability grades. Section IV delves into my use of the ISTEP scores in this analysis.

i. Item Response Theory

ISTEP tests utilize a theory of psychometric evaluation of test results known as item response theory (IRT). IRT treats each question, or item, in a test as independent from one another. In this way, the item is the object of analysis, not the entire test ("Item Response Theory," 2019). Additionally, IRT assumes that each individual taking the exam has a certain level of the latent ability being tested by the item (Primi et. al., 2016). For each item on the exam at a given the

[2] ISTEP is referred to as ISTEP+ once the Science section is added. However, third grades don't sit for the science section so the "+" is left off of references to it in this paper. All of the information regarding the ISTEP test form itself is identical with and without the "+".

ability level of the examinee, there exists a certain probability that the examinee will correctly answer the item (Primi et. al., 2016). A link can then be established between the underlying characteristics on the ability scale and the response of the particular individual (Primi et. al., 2016). The graph below shows the S-shaped curve which "describes the probability of a correct response to an item as a function of possessed ability" known here as theta (Primi et. al., 2016).



This graph from Primi et. al. (2016) shows that when the difficulty of the item, shown by line *b*, functions at a possessed ability level of one, then the probability of responding correctly to the item is 50%. Line *a* shows that at this difficulty level, the item can discriminate very well between low- and high- ability examinees; the probability of a correct response changes quickly with a change in ability level. This particular model of IRT is known as the two-parameter logistic model, the two parameters being difficulty and discrimination.

Because of this factor model used by IRT, the statistical probability of responding correctly to an item can negatively or positively impact a student's score when that score is simply the raw number of items rightly answered. Pattern scoring used in conjunction with the IRT allows item characteristics like difficulty to be taken into account when assigning ability scores. For a 50

question test, if two students answer all items correctly but one answers the 25 most difficult questions and one answers the 25 easiest ones, then the student who answered the more difficult questions will receive a higher ability score ("Indiana Department of Education "ISTEP+"", 2017, p. 2).

ii. Criterion Referenced Scores

The ISTEP test is also a criterion-referenced test. The items on the test aim to assess a student's abilities with respect to a specific criterion which is, in this case, outlined by the Indiana State Board of Education (ISBE) in the form of Indiana Academic Standards. A student's ISTEP test results are not to be compared to student test results nationwide, but to a "cut score, defined by educators, and based on Indiana Academic Standards" (Indiana Department of Education "ISTEP+", 2017, p. 1). Essentially, the test results are comparable only between Indiana students who took that particular test. The cut score divides student performance between Did Not Pass, Pass, and Pass+ (p. 4). Criterion-referenced tests like ISTEP also identify a student's weaknesses and strength so that the student can be more appropriately educated (Indiana Department of Education "ISTEP+", 2017, p. 1).

iii. Scale Scores

The ability scoring used by IRT in conjunction with criterion-referencing allows ISTEP to transform a student's ISTEP test scores on a general ability scale, making the score more meaningful to the particular student and in comparing a student's progress from year to year. The ISTEP tests use two scales, one for ELA and one for Math, which move vertically. These scales are kept constant from grade three to eight, although grade three typically has a lower range. The ISBE defines the lowest obtainable scale score and the highest for each grade's scale. The cut score determined by the ISBE is on the scale and acts as a reference for where each student's proficiency

level lies (Indiana Department of Education "ISTEP+", 2017, p. 3-4). It is each school's third grade average ELA scale score and average Math scale score which I use for this paper. My methods are elaborated upon in part A of Section III.

Part D: Test Scores in Indiana

It should be noted that on top of the state grading systems, there is a federal grading system that is used to assess the performance of schools. During the NCLB, Indiana received a waiver from the federal government that exempted them from having a state accountability system and a federal system (Grew and Sheldrake, 2013, p. 8). This meant that federal and state allocations and actions were determined solely upon the accountability model which Indiana created for itself. Therefore, the federal and state government used the state's accountability score to determine respective resource allocation and consequences. When the ESSA was signed into law, the federal government required Indiana to submit a new plan in accordance with the new accountability guidelines (A. Whelan, personal communication, December 15, 2015).

The result is such that Indiana currently has two sets of accountability grades, one in compliance with ESSA and one based on their own model which was used singly until the 2017-2018 school year. Beginning in the 2018-2019 school year and continuing "into the foreseeable future," Indiana will maintain two different grading systems for the state and the federal government (M. Paino, personal communication, January 24, 2018). State allocations and consequences will be determined by the state score based on the state's system, and federal allocations and consequences will be determined by the system submitted to and approved by the federal DOE's resultant score.

The method by which the Indiana Department of Education calculates school accountability scores for elementary-middle schools (EMS) in Indiana, the class of schools I am

working with for this thesis, is broken down into two domains: growth and performance. Because I am focused simply on the third grade for reasons discussed in part B of Section II, the entire accountability score is awarded to the third grade by simply the Performance Score. The plan submitted to the federal DOE on September 18, 2017, which is to be used by the federal government to grade Indiana schools as mentioned above, added two indicators: a School Quality/Student Success Indicator measured by levels of chronic absenteeism, and an English Language Proficiency Indicator ("Accountability," n.d.).

The IDOE is due to submit amendments to their ESSA-compliant accountability scoring system to the federal government in 2019. The only new indicator being proposed for EMS is a Closing Achievement Gaps Indicator, which measures "subgroup proficiency on the English/language arts & mathematics assessments" and "compares subgroup achievement to [a] statewide long-term goal for each subject area and subgroup to determine overall performance" ("ESSA Amendment Accountability Summary," n.d., p. 14). This new indicator is one purely for disadvantaged subgroups and how they performed compared to state-determined benchmarks to see how well the school is helping these students. However, these changes apply only to the federal grading system. As I mentioned in part B(ii) of Section I, I will be focused on the state accountability system and thus these do not apply to my models. Below I have explained the component of the state-assigned accountability grades relevant to the third grade.

Part E: The Relevant Accountability Scoring Calculation—Performance Domain

If the IDOE evaluated schools on a grade-to-grade basis, then the third grade's score would come solely from the Performance Domain. This is the domain for which I will dedicate an explanation for how it is calculated. There are two components to the Performance domain: a Math Indicator and a English Language Arts (ELA) Indicator. Both indicators are calculated by

multiplying the percentage of students who passed each section's respective assessment by the number of students who completed the exam and received a valid result for each respective assessment. The first number is the "pass rate," and the second number is the "participation rate." If the participation rate is greater than 95% for an indicator, than it becomes a multiplier of one[3].

The Math Component and the ELA Component are averaged together for the final score of the Performance domain; in the case of the third grade, the Performance domain's score is equal to its entire final score which translates directly to an accountability grade. It is relatively clear from even just this short description that there is no rigorous theory behind the grade by which a school is assessed year after year. No mathematical attempts are made to account for anything; the passage rates of the ISTEP test and the percentage of the grade at each school which participates in the test are simply calculated to form the accountability grade. It is obvious that more than these two things reveal the performance of a school, and that a myriad of factors contribute to a student's performance on the test itself.

III. DATA AND METHODS

> "The basic problem is that schools can be no better than the residents of the school district." –Phillip Hermes, WSJ March 26, 2019

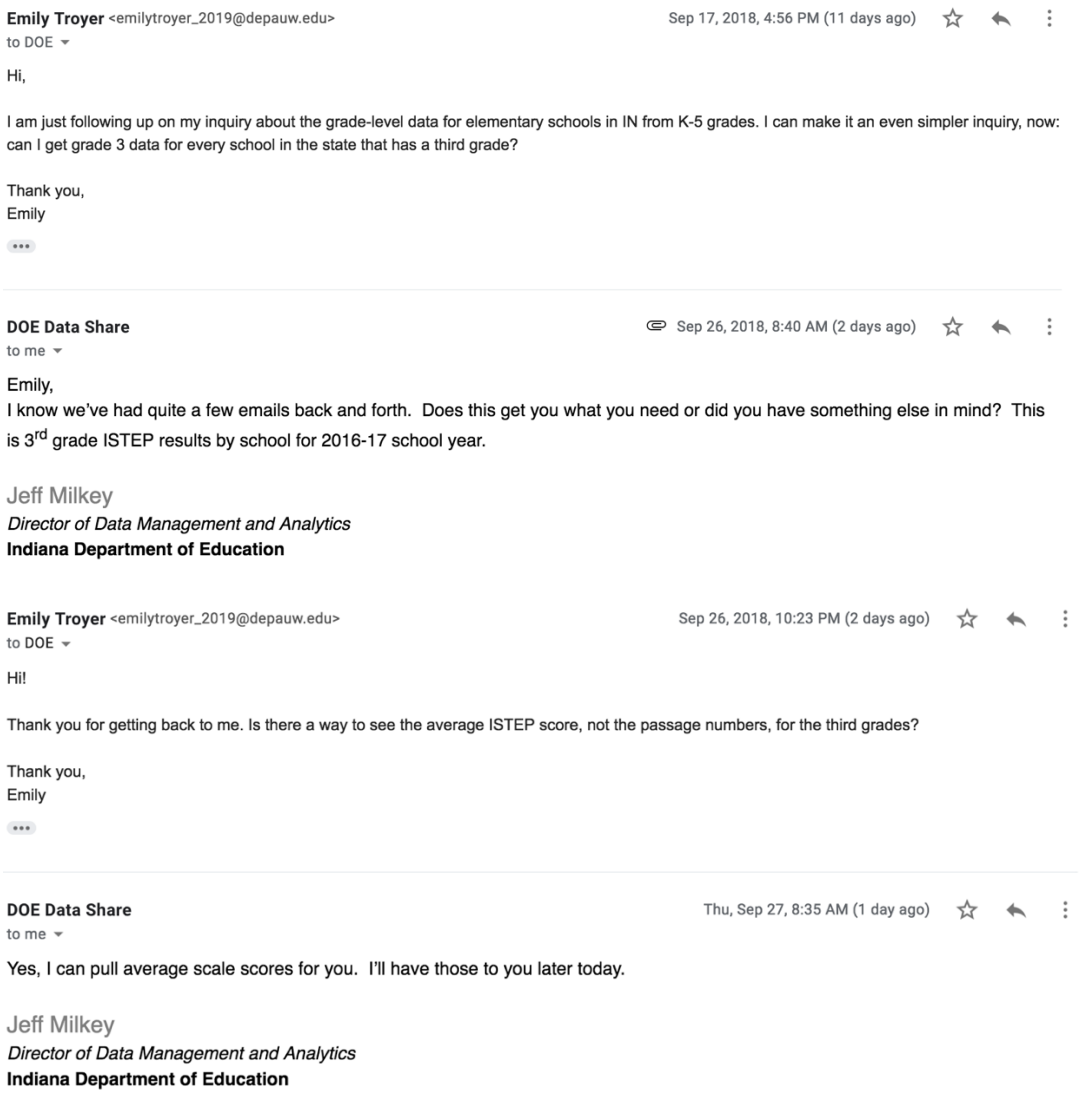
Part A: Obtaining Grade 3 ISTEP Scaled Scores

A weighted average of the two components of the ISTEP exam—English/Language Arts (ELA) and Math—is the dependent variable in this regression analysis. From the "Accountability" page on the IDOE website, I obtained the scaled scores of all reporting third grades in the state of

[3] Example: an EMS had a passing rate of 85% for Math and a participation rate of 91%, the Math Component of the performance domain would be 0.85*0.91 = 0.7735 points.

Indiana. On that page the "Find School and Corporation Data Reports" page is nested. This page contains links to downloadable data files that contain various breakdowns of ISTEP scores for schools and corporations in the state of Indiana. Unfortunately, average ISTEP scores for uniquely grade three were not available online. I then emailed the listed contact email, datashare@doe.in.gov, in hopes of being able to obtain the grade three ISTEP score data for each school in the state of Indiana. The IDOE sent me the third grade ISTEP scores for 2015-2016. Screenshots of our conversation are pictured below in Figure 1.

Figure 1: Correspondence with the IDOE regarding ISTEP scores needed.



Emily Troyer <emilytroyer_2019@depauw.edu> Sep 17, 2018, 4:56 PM (11 days ago)
to DOE

Hi,

I am just following up on my inquiry about the grade-level data for elementary schools in IN from K-5 grades. I can make it an even simpler inquiry, now: can I get grade 3 data for every school in the state that has a third grade?

Thank you,
Emily

DOE Data Share Sep 26, 2018, 8:40 AM (2 days ago)
to me

Emily,
I know we've had quite a few emails back and forth. Does this get you what you need or did you have something else in mind? This is 3rd grade ISTEP results by school for 2016-17 school year.

Jeff Milkey
Director of Data Management and Analytics
Indiana Department of Education

Emily Troyer <emilytroyer_2019@depauw.edu> Sep 26, 2018, 10:23 PM (2 days ago)
to DOE

Hi!

Thank you for getting back to me. Is there a way to see the average ISTEP score, not the passage numbers, for the third grades?

Thank you,
Emily

DOE Data Share Thu, Sep 27, 8:35 AM (1 day ago)
to me

Yes, I can pull average scale scores for you. I'll have those to you later today.

Jeff Milkey
Director of Data Management and Analytics
Indiana Department of Education

Part B: Obtaining School Accountability Scores

In order to obtain the accountability scores administered by the Indiana Department of Education (IDOE), I navigated to the IDOE's Office of Accountability web page. A small list on the right hand side of this page links to the IDOE's "Find School and Corporation Data Reports" web page. Towards the bottom of the available data was a series of the state's A-F grade results for schools and corporations by year. I was then able to easily pull the Excel files for school-level data for the school years 2015-2016 and 2016-2017. These initial files contained accountability scores for every school in the state of Indiana that takes the ISTEP test. In the end, I used only the data from 2015-2016 because that is the most recent year for which the independent variables used in this thesis were available at the time I gathered this data.

Part C: Obtaining Independent Variables

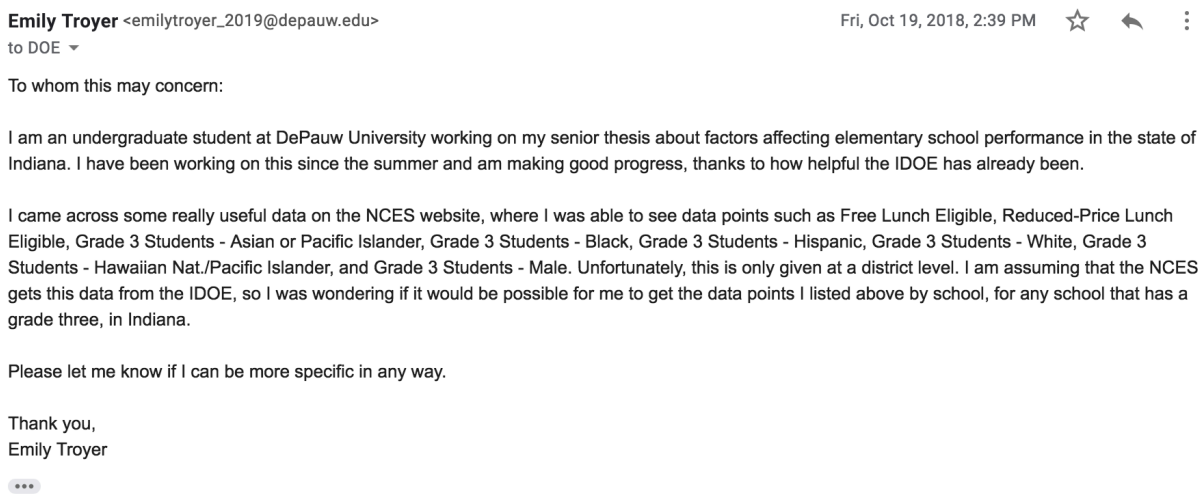
The purpose of this regression is to hold constant factors outside the school's control affect third grade ISTEP scores in Indiana. Possible factors include but are not limited to socioeconomic status, unemployment rates in respective school districts, race, gender, parent presence in education, and marital status of parents. In fact, there is a nearly limitless count of variables I would ideally like include in this study, many of which are difficult to measure like intelligence or raw ability. The only included independent variables in this thesis are ones that I could obtain from the IDOE.

The first variable I set out to find was the number of third grade Hoosiers per school who were eligible for a free lunch. This would allow me to have a proxy for the particular socioeconomic status of the student's family, and act as a measure of the stress the student may be under at home. When I searched this query, I found that the National Center for Education Statistics

(NCES) reports the number of students eligible for a free or a reduced price meal at each school. This value is based on the student's family's income and poverty thresholds set by the federal government each year. See Appendix B for specific guidelines for the 2015-2016 school year.

Not only does NCES have the number of students eligible to receive free and reduced price lunches, but they also list race and ethnicity demographics and gender demographics for Indiana schools during the 2015-2016 school year. All of this information is accessible at a school level, not at the third-grade level. Assuming that the NCES likely receives this information from each state, I reached back out to the IDOE to get this information broken down for only the third grades in Indiana. The IDOE Data Share was again helpful, and sent this data in Excel files at the third grade level for reporting Indiana schools. It is clear that there was some confusion at first, but eventually I did get the data that I needed. See Figure 2 below for our correspondence.

Figure 2: Correspondence with the IDOE regarding independent variables needed.



Emily Troyer <emilytroyer_2019@depauw.edu> Fri, Oct 19, 2018, 2:39 PM

to DOE

To whom this may concern:

I am an undergraduate student at DePauw University working on my senior thesis about factors affecting elementary school performance in the state of Indiana. I have been working on this since the summer and am making good progress, thanks to how helpful the IDOE has already been.

I came across some really useful data on the NCES website, where I was able to see data points such as Free Lunch Eligible, Reduced-Price Lunch Eligible, Grade 3 Students - Asian or Pacific Islander, Grade 3 Students - Black, Grade 3 Students - Hispanic, Grade 3 Students - White, Grade 3 Students - Hawaiian Nat./Pacific Islander, and Grade 3 Students - Male. Unfortunately, this is only given at a district level. I am assuming that the NCES gets this data from the IDOE, so I was wondering if it would be possible for me to get the data points I listed above by school, for any school that has a grade three, in Indiana.

Please let me know if I can be more specific in any way.

Thank you,
Emily Troyer

DOE Data Share <datashare@doe.in.gov> Mon, Oct 22, 2018, 2:19 PM ☆
to me

Hi Emily,
Are you requesting the following data points. Please correct me if I have misunderstood your request.

1. Free and Reduced Lunch data for Grade 3 Students - Asian or Pacific Islander
2. Free and Reduced Lunch data for Grade 3 Students - Black,
3. Free and Reduced Lunch data for Grade 3 Students - Hispanic,
4. Free and Reduced Lunch data for Grade 3 Students - White,
5. Free and Reduced Lunch data for Grade 3 Students - Hawaiian Nat./Pacific Islander, and
6. Free and Reduced Lunch data for Grade 3 Students - Male.

Thanks,

Emily Troyer <emilytroyer_2019@depauw.edu> Mon, Oct 22, 2018, 2:25 PM ☆ ↩ ⋮
to DOE

Hi Hammad,

Sorry for the confusion--let me clarify. What I am requesting is, at a school level, so for each school in the state of Indiana with a third grade, the following data points:

1. Free and Reduced Lunch data for Grade 3 Students
2. Grade 3 Students - Black,
3. Grade 3 Students - Asian or Pacific Islander
4. Grade 3 Students - Hispanic,
5. Grade 3 Students - White,
6. Grade 3 Students - Hawaiian Nat./Pacific Islander, and
7. Grade 3 Students - Male.

Thank you,
Emily

DOE Data Share <datashare@doe.in.gov> Oct 22, 2018, 2:31 PM ☆ ↩ ⋮
to me

Hi Emily,
So you just need the student enrolment broken down by ethnicity and gender for grade 3 students. The most recent data available is 2017-18.

Let me know.

Thanks,

Emily Troyer <emilytroyer_2019@depauw.edu> Oct 22, 2018, 2:42 PM ☆ ↩ ⋮
to DOE

Hi Hammad,

Yes, on a school level not a district or state level for each point, as well as the data points of Free Lunch Eligibility for grade 3 students at each school and Reduced Lunch Eligibility for grade 3 students at each school.

If I could get this information for 2015-2016 and 2016-2017 that would be great.

Thank you,
Emily

DOE Data Share <datashare@doe.in.gov> Oct 22, 2018, 3:00 PM

to me

Hi Emily,

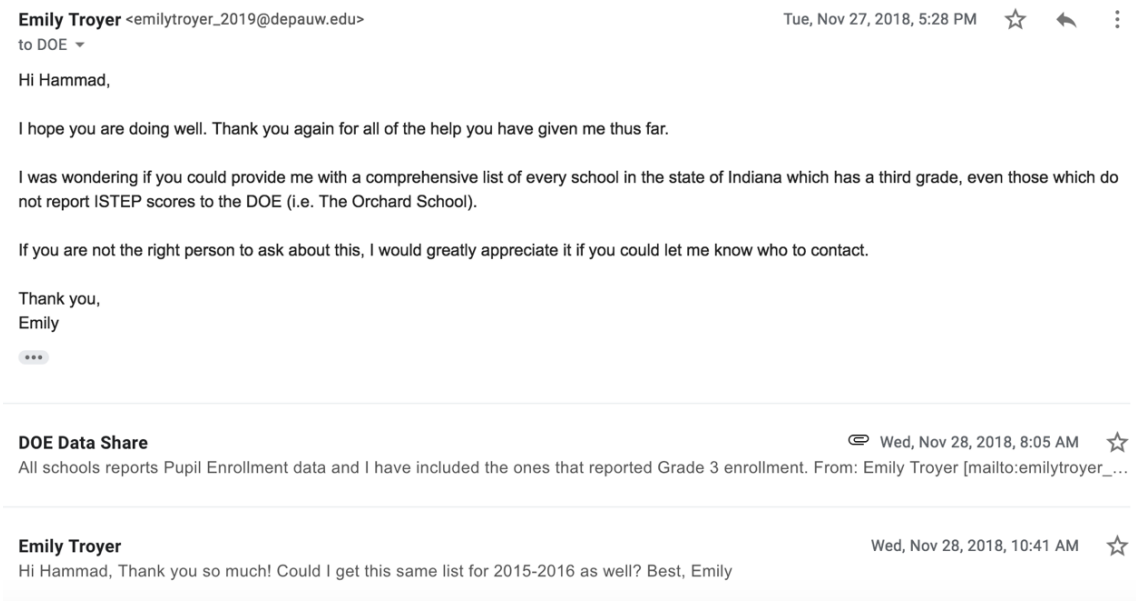
Please see attached as a response to your request. Let me know if you have any questions or concerns with the attached data. If you have a moment, we invite you to take a quick survey on the service you received from the Data Request Team: Customer Survey

Thanks,

The last explanatory variable I sought was the number of third graders at each school who are English Language Learners (ELLs). This is one of the ESSA's target groups and is a subgroup delineated by the IDOE (McCormick, n.d., p. 26). It is clear why they are a target group: students who are not already fluent or proficient in English before coming to the school will struggle academically to adjust. Schools can and should help teach students English, but ought not to be held entirely accountable for the disadvantage an ELL student is because of his or her lack of English skills. The IDOE Data Share was once again able to send this file upon my request.

I also needed the total population of each school's third grade so that I could convert the demographics from raw numbers into percentages of the school's population. Figure 3 below shows my correspondence with the IDOE Data Share to obtain the total number of third graders at each school. The IDOE sent me an Excel file with the total populations of third graders at each school.

Figure 3: Correspondence with the IDOE regarding third grade population counts needed.



Emily Troyer <emilytroyer_2019@depauw.edu> Tue, Nov 27, 2018, 5:28 PM
to DOE

Hi Hammad,

I hope you are doing well. Thank you again for all of the help you have given me thus far.

I was wondering if you could provide me with a comprehensive list of every school in the state of Indiana which has a third grade, even those which do not report ISTEP scores to the DOE (i.e. The Orchard School).

If you are not the right person to ask about this, I would greatly appreciate it if you could let me know who to contact.

Thank you,
Emily

DOE Data Share Wed, Nov 28, 2018, 8:05 AM
All schools reports Pupil Enrollment data and I have included the ones that reported Grade 3 enrollment. From: Emily Troyer [mailto:emilytroyer_...

Emily Troyer Wed, Nov 28, 2018, 10:41 AM
Hi Hammad, Thank you so much! Could I get this same list for 2015-2016 as well? Best, Emily

Holding all of these variables constant—gender, race, ethnicity, location, socioeconomic status, and ELL status—I hoped would help predict the average ISTEP score a school's third grade would receive, and therefore a more accurate measure of the efficacy of the school.

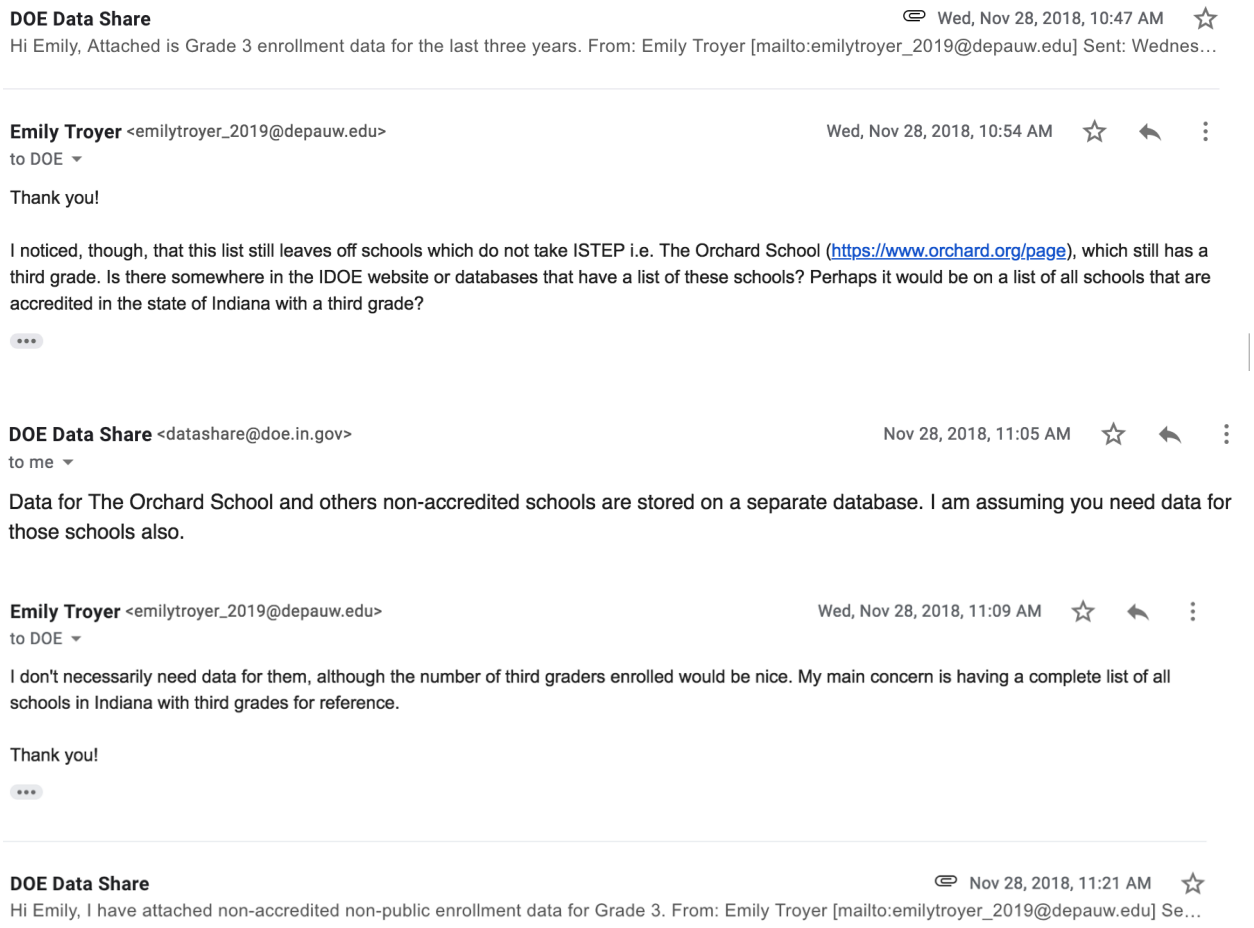
Part D: A Brief Note on School Accreditation in Indiana

Because I wanted to include every school in Indiana which has a third grade, I was concerned when schools containing third grades that I knew of from personal experience were not included in the information I had received. This could skew the results because I saw that, at minimum, the data was missing two of the most well-regarded private grade schools in Indianapolis; if the analysis is missing some high-performing schools, then effects of external factors on the ISTEP scores may be over- or under- credited.

After receiving the messages from the IDOE Data Share shown below in Figure 4, I investigated what it means to be accredited or not-accredited in the state of Indiana. Per the IDOE's "School Accreditation" subpage of their "Accountability" webpage, full accreditation at the school

level comes when a school receives either an A, B, or C from the IDOE's Office of Accountability and follows all state and federal laws. If a school receives either one D or as many as three consecutive F's, then it is assigned a provisional accreditation. The state board assigns a subcategory of provisional accreditation called "Provisional: Legal Standards" if the school has not fulfilled the requirements for provisional but has broken a law. Finally, a school receives a probationary accreditation if it has earned at least four consecutive F's ("School Accreditation," 2017). This is important in its own right because to whom and how much state funding is allocated depends upon a school's accreditation status (Harsanyi, M., personal communication, January 31, 2019).

Figure 4: Correspondence with the IDOE regarding missing schools.



Because some of the wealthy, private schools in Indianapolis do not participate in ISTEP testing, they are not eligible for state accreditation and are not given accountability scores. Without ISTEP scores, the school's third grade cannot have a dependent variable by my calculation methods. Therefore, this analysis will not be able to include non-state accredited schools, and the risk of over- or under- crediting the effects of the included independent variables on average ISTEP scores persists. This is a limitation of my study.

Part E: Cleaning the Data

Each of these variables, both dependent and independent, were sent in separate Excel files. In order to use this data to run a regression, I needed to first merge them. Each set of data came with a corporation ID and a school ID. To identify each individual school uniquely so the data sets could be merged, I concatenated the corporation and school ID's into a *CorpSchoolID* variable for each sets of imported data in their respective Excel files. Since some schools in different corporations have the same name, this exercise was extremely useful. I also made sure that each data set's first row was the label of each variable in the set.

I began to merge the data sets in Excel by copying the information from the third grade ISTEP scores XLS file into a blank Excel workbook. Then, I copied the information from the accountability file into that same workbook. Using Excel's VLOOKUP function, I was then able to assign a third grade ISTEP score to the accountability score by matching the accountability score data set's *CorpSchoolID* number with that of the third grade ISTEP score data set. After doing this, however, I realized that using Excel to merge multiple files would be tricky and risky. Because of the this, I stopped the merging process in Excel.

Using STATA, I converted each of the original files for the independent variables, including the number of pupils enrolled in third grades at schools, into .dta files. I then created a

master .dta file populated initially with only the ITSEP score data. I then merged each of the .dta files for the IDOE accountability grade, pupil enrollment, locale, ELLs, race and ethnicity, gender, students who received a free lunch, and students who got a reduced price lunch with the master file. When STATA merges data sets, observations with the same unique identifier in the separate sets, in this case *CorpSchoolID*, become the same observation.

The variable *avg_score_w* was the first that I created in STATA. It takes a weighted average of each school's third grade's average scale score of ELA and of Math; this is the dependent variable of this thesis. I use a weighted average because in some instances, different numbers of students sit for the ELA test and the Math test. Furthermore, I want a single indicator of performance and if there is more information on one measure than I want to weight that measure appropriately to reflect its presence in the single indicator.

I do recognize that even taking a weighted average of ELA and Math scores presents some issues. One issue that arises is that there may have been different scales for ELA than there were for math. Another issue is that perhaps a school is stronger in ELA instruction than in Math instruction; this average would mute that difference in student performance. Furthermore, it is possible that some of the included independent variables affect ELA performance more than Math or vice versa. This average score would again muddy that effect. In a perfect world, these would not be ignored. However, this world is not perfect and in this thesis I am attempting to check for any effect these outside factors have on the general *avg_score_w*.

Next, I dropped all of the observations which did not have *avg_score_w* variables and any observation where ten or fewer students sat for either portion of the ISTEP. This brought me to 1,280 observations. It should be noted that the schools which were removed included independent private schools, schools transitioning into middle schools at the time of the 2015-2016 ISTEP test,

and youth opportunity centers which do not have a formal school infrastructure but are included in a school corporation. While the analysis will miss the insight from these unique institutions, enormous variability occurs in a sample size of one. Furthermore, if no one sat for one portion of the test (ELA or Math), then the *avg_score_w* for that school is not an average of the school's Math and ELA scores and is therefore incomparable to other schools' *avg_score_w*.

After dropping these extreme observations, I generated percentage versions of the independent variables that will be used in the regression. These variables need to be in percentage form so that the independent variable represents the proportion of students in the school who have each characteristic. This will allow for an easier interpretation of the regression results. First I made sure that the imported data had zeroes in place of all the null cells. To create the proportions, I used STATA to take the number of ELLs, the number of each race and ethnicity, the number of males, the number of students who were eligible for a free lunch, and the number of students who were eligible for a reduced price lunch in each school divided by the observation's corresponding pupil enrollment.

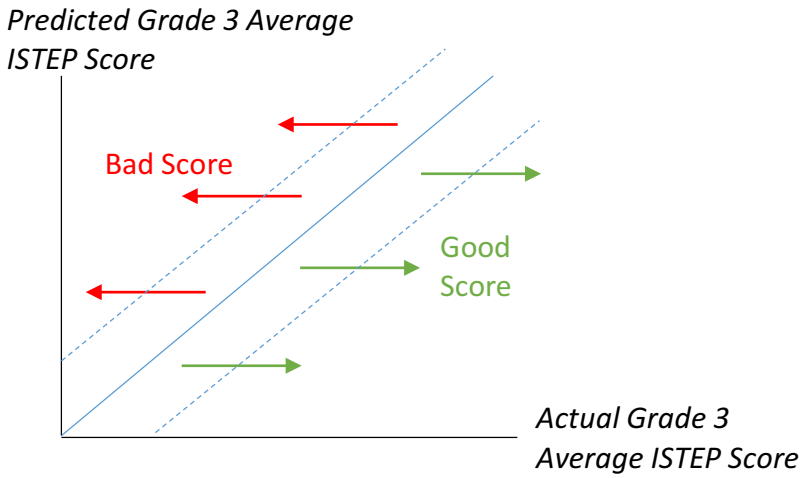
Part F: Goal of this Evaluation

As Section II part A established, third grade is a crucial time in a child's education. For this reason, I first create an accountability score for the third grade according to the IDOE's methods for assigning scores at a school level. I then create an alternate grading system which relies on three distinctions: "Exceeds Target," "On Target," and "Does Not Meet Target" in order to identify exceptionally performing schools. To do both of these tasks, I use regression to control factors that contribute to the ISTEP score which are outside of a school's control. The long model used in this analysis is below:

$$AvgScore16 = \beta_0 + \beta_1 Male + \beta_2 AmericanIndian + \beta_3 Asian + \beta_4 Black + \beta_5 Multiracial + \beta_6 HawaiianPacificIslander + \beta_7 Hispanic + \beta_8 ReducedPriceLunch + \beta_9 FreeLunch + \beta_{10} ELL + \varepsilon_i$$

After obtaining predicted ISTEP scores holding certain demographic factors constant, I plot the predicted and real average ISTEP scores on a graph with a forty-five-degree line; the actual grade three average ISTEP score as a function of predicted average score. Figure 5 below demonstrates this idea.

Figure 5: Outline of My Proposed Accountability Grading System



Scores to the right of the solid line indicate that the actual average ISTEP score is better than what I predicted it would be; the school is doing better than it "should" be doing. However, if actual scores fall to the left of the solid line, then the school is doing worse than it "should" be doing. Schools on either side of the dashed lines are the extreme performers and are denoted by the descriptions "Exceeds Target" and "Does Not Meet Target." One could also use the steps provided in Section III part A to imagine the grade the IDOE would assign to each third grade given the adjusted and unadjusted ISTEP scores. This is something that I do and expand upon in the next section.

IV. MODELS AND RESULTS

"On [the principal's] worst days, she told me once, "I don't think people really think it's possible," referring to turning around [Atlanta] neighborhood schools like Peyton Forest... the profession… seems to require bifocal vision: an ability to see the dispiriting big picture, but also an ability to see the child close at hand" –Sara Mosle, *New York Times Magazine*, September 6, 2018

Part A: Regression Models & Revised Grading System

Figure 6, part A below shows the summary statistics of the sample used in the regressions I run. One of the most interesting statistics is that the average percentage of students on free lunch in a given school is 44%. The variable *free_lunch* acts as a proxy for socioeconomic status so it is somewhat surprising that the average is so high. It should be noted, though, that the standard deviation of this is +/- 26 percentage points, so if the sample is normally distributed then 68% of the sample will have 18% to 70% of students on free lunches. The variation here is enormous, and helps to explain why this variable is so impactful in the regression models.

It should also be noted that the average number of males per school in the sample is 51% +/- 8 percentage points. This seems to indicate that gender is relatively equal for most observations. It is clear that there are some all-male schools in the data set because the maximum percentage of males is 100%. The minimum is not exactly zero, so it is unclear whether or not there are any all-female schools in the data set.

The racial composition of the sample as described in Figure 6 part A is interesting as well. The averages for *amerindians, asian,* and *pacisl* are almost identical to the proportions of all three races in Indiana as a whole. The other race variables—*hispanic*, *black*, and *mraces*—over-represent their groups compared to the proportions of these races in Indiana as a whole. However, their standard deviations are so large that they capture Indiana's demographic makeup ("Census

estimates…", 2014). This indicates that high levels of variation between these racial groups between schools exist in my sample.

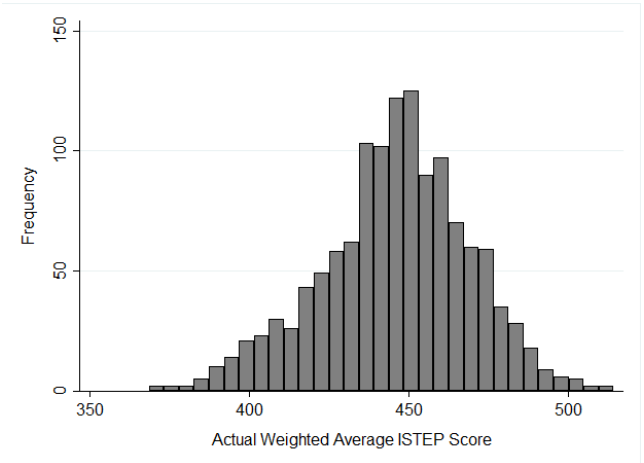
The histogram in part B of Figure 6 below shows that the average weighted ISTEP scores in this sample are skewed-left by a minute amount. Generally, most third grades have an actual weighted ISTEP score of 445 to 455. Neither tail is particularly dramatic, so the dependent variable is not riddled with outliers. This is good for predicting purposes, as it can be difficult to accurately predict outcomes when the sample is full of outliers.

Figure 6: Summary Statistics

A. Summary Statistics Table

Variable	Obs	Mean	Std. Dev.	Min	Max
free_lunch	1,280	.439971	.2586279	0	1
rp_lunch	1,280	.069782	.0503697	0	.4
males	1,280	.5091261	.0776212	.0147059	1
ell	1,280	.0613273	.1044926	0	.765625
amerindians	1,280	.0018668	.0065516	0	.0645161
asians	1,280	.0163103	.0390923	0	.4270833
blacks	1,280	.1197421	.208188	0	1
hispanics	1,280	.1187636	.1543514	0	1
mraces	1,280	.050097	.0470202	0	.3333333
pacisl	1,280	.0007966	.0047729	0	.0625

B. Histogram of Actual Weighted Average ISTEP Scores, 2015-2016



Below is the regression output for the models I am using. I use three models to demonstrate the power of controlling for specific characteristics when considering third grade student achievement in Indiana.

Figure 7: Regression Output Table

Regressing Third Grade Weighted Average ISTEP Scores on Demographic Characteristics

	Model 1	Model 2	Model 3
free_lunch	-64.4729***	-61.5658***	-47.0699***
	(2.1e+00)	(2.3e+00)	(2.8e+00)
rp_lunch		3.3008	-18.5864*
		(1.0e+01)	(9.0e+00)
males		-17.6432**	-16.2060*
		(6.8e+00)	(7.1e+00)
ell		-16.6700***	-41.1697***
		(4.7e+00)	(7.8e+00)
amerindians			-39.5400
			(6.1e+01)
asians			73.1053***
			(1.2e+01)
blacks			-25.9401***
			(3.6e+00)
hispanics			11.8834*
			(5.1e+00)
mraces			-13.7215
			(1.2e+01)
pacisl			-56.8799
			(1.2e+02)
_cons	474.1784***	482.6739***	479.9033***
	(1.0e+00)	(3.7e+00)	(3.8e+00)
N	1280	1280	1280
R^2	0.514	0.522	0.572
adj. R^2	0.514	0.521	0.568
F	961.1328	266.9050	146.6332

Robust standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Model 1, my short regression, is used simply to show the power of holding something constant. Even adding in just one powerful control variable—*free_lunch,* a proxy for socioeconomic status—explains over half of the variability in average weighted ISTEP scores for third grades across the state of Indiana. The coefficient estimate for the effect of *free_lunch* on *avg_score_w* predicts that if *free_lunch* increases by one percentage point, then the *avg_score_w* will increase by -64.4729 +/- 0.021 percentage points.

Model 2 introduces another socioeconomic proxy, *rp_lunch* as well as an English-language learner status variable, *ell*. There is correlation between *free_lunch* and *rp_lunch*, but not multicollinearity. Because they are correlated explanatory variables, it is important that both are included to avoid unnecessary omitted variable bias. Intuitively, it is correct that *rp_lunch* affects *avg_score_w* with less magnitude than *free_lunch*. Students who qualify for reduced-price meals are slightly better off socioeconomically than students who qualify for free meals, and would therefore have a slightly more learning-conducive home environment. Adding more explanatory variables muted the effect of *free_lunch*, but not by much. This confirms the importance of socioeconomic status in explaining third grade performance on test scores.

I choose to leave race out in this model because there are some technical issues with racial definitions at the IDOE, discussed more in-depth below, and wanted to see the effects of additional variables without the complications of this error. The *ell* variable includes students who do not natively speak English of any race. Overall, Model 2 explains 52.1% of the variation in *avg_score_w* after adjusting for the increase in variables. While this is a marginal increase in explanation, it is an increase nonetheless.

Model 3 explains the most variation in *avg_score_w*—56.8% after adjusting for the increase in independent variables. In this model, *female* is the gender base case, so the *males*

variable is compared to females. According to a potentially groundbreaking working paper by Reardon, Fahle, Kalogrides, Podolsky, and Zárate (2018), third grade through eighth grade males and females in 10,000 school districts across the nation do the same on average in math; some variation in favor of each gender occurs between school districts. Male-dominant achievement gaps in math occur, however, when the school district is socioeconomically advantaged. Conversely, poor girls of color do better than poor boys of color in math and ELA.

Their study using over 260 million observations, introduces ambiguity into the literature which overwhelmingly shows males excelling over females on math portions of standardized tests. Given the pre-existing literature and this working paper, it makes sense for my *males* coefficient estimate to be ambiguous; the dependent variable is a weighted average of both the math and ELA components of the ISTEP test, so the gender effect is muddled. Furthermore, the percentage of observations in my data set where less than half of third graders in the school are eligible for free lunch is 60%. This means that in 40% of the schools in this data set, half or more of their third graders are eligible for free lunch. This indicates a roughly even mix of predominantly high- and low-socioeconomic status schools, which further supports an ambiguous *males* estimated effect[4].

White is the race and ethnicity base case; all of the race and ethnicity variables are compared to white. The racial variables *amerindians, mraces,* and *pacisl* are all statistically insignificant, which makes sense considering the fact that the average proportions of these racial groups per third grade in this data set are 0.19%, 5.01%, and 0.08% respectively.

The IDOE's classification of race and ethnicity presents some issues which surface in Model 3 which stem from the fact that it considers Hispanic, an ethnicity, to be a race. Organizing

[4] For more justification for the viability of my data set, see my regression tables for gender's effect on both math and ELA ISTEP scores in Appendix D.

race variables this way ignores the existence of white Hispanics because in this sample, a student can either be exclusively white or Hispanic. This leaves me with a positive yet significant *hispanic* coefficient estimate, indicating the confounding which results from some Hispanics being white. To be clear, I am not arguing that being Hispanic contributes to a higher ISTEP score. What happens here is that when *hispanic* is compared to the base case *white*, it performs positively albeit with a small effect. This quirk is unimportant to my thesis, however, because the point is that race and ethnicity is being controlled for when predicting average weighted ISTEP scores for third grades in Indiana.

Another thing which requires some attention is the error term. It is the unknown error in the predicted results and includes everything not accounted for in the regression that causes volatility in the estimations—omitted explanatory variables, measurement error, and most importantly for this study, the level of luck on guesses on test questions and the conditions of the student and environment on the day of the 2015-2016 ISTEP test. Because my sample is not student-level, the level at which test scores are generated, the data generation process for the sample entails third-grade children taking the test, and then taking the weighted average of both sections of the third grade's test scores.

The issue lies in the nature of test scores: they vary each time the test is taken. A student may take it once and get one grade, then take the same test again and get different; he certainly guessed on some questions and could get different luck levels on those guesses. Additionally, on the day of either test he could have been sick, slept little, had extreme nerves, or experienced a traumatic event. These all create variation in the test score he receives each time he takes the test. do not have scores of individual students, just averages of the school at a grade-level; Because I only have averages at a grade-level, I am not just working with his deviating scores. I am dealing

with the standard deviation of each student's individual score at a given school because *avg_score_w* is the average of multiple kids; then, the deviation of the average is its standard deviation, which is unknown so we estimate it with the standard error, over the square root of the number of students in the school's test-taking population. This creates a persistent problem in all three of my models: heteroskedasticity.

Heteroskedasticity occurs when the error terms—the unknown error in the predicted results because of measurement errors, omitted variables, and random chance—are not identically distributed for each observation. In this particular case, heteroskedasticity occurs largely due to the fact that each observation does not have the same size of school. The variation in test-taker sizes between schools is massive in this sample, ranging from 9 to 242 students in a given third grade. The estimated standard error (SE) reported by the regression results estimates a single error term distribution for all observations. This is problematic because there is a unique error term distribution for each observation stemming from the range in the number of students for each observation; those with larger numbers of test-takers have less fluctuation. I account for heteroskedasticity via a robust standard error (RSE). RSE's estimate the error term distribution based on each observation, instead of trying to estimate the entire sample's error term distribution.

An additional source of error in this model is the base measure of the accountability grade: the ISTEP scores themselves. Because standardized tests like ISTEP are subject to conditions on the day of the test, ranging from student sleep levels to the time of day the test was administered, the error terms in this sample have a lot of variation. Since the outcomes of the test are massively random, the coefficient estimates in all models are volatile. I am not concerned with the specific effect of each independent variable, as I acknowledge that many more independent variables are needed to fully explain the ISTEP scores and adding more explanatory variables would alter the

coefficient estimates. Therefore, this is low priority error source for my analysis. If policymakers wanted to analyze what parts of student background affect their testing outcomes the most, however, this error source would be a major issue.

Part B: Results

Figure 8: Third Grade Accountability vs. All School Accountability

Avg. Weighted Third Grade ISTEP Score	IDOE School Grade	Grade 3 Grade
472.96	A	D
460.77	A	D
453.74	A	D
460.07	A	D
466.65	A	D
462.22	A	D
457.44	A	D
460.37	A	D
457.65	B	F
421.76	B	F
398.27	B	F
415.65	A	F

In Figure 8 above, I assign an accountability grade for the third grade based on the same methods employed by the IDOE to calculate a school's accountability grade given the same information. Essentially, my grade excludes the Growth domain because third grade is the first year that students take the ISTEP. The grade is simply calculated on the raw Performance domain. For more information on the grading systems, refer to part E of Section II.

There are 193 elementary schools[5] in the state of Indiana where my score for the third grade is three or four letter grades worse than the grade given by the state. This indicates a huge deviation between the performance and growth of students in third and fourth grade at each particular school. There are 24 elementary schools in Indiana in which the IDOE assigns a worse grade than my grade, some of them with more significant impact than others such. One observation has an IDOE accountability score of an F, but my third grade score is a D. This is a critical difference. Nuance, then, is crucial in assessing the performance of a school. Different teachers and different curriculum exist for different grade levels, and some may be more effective than others.

For the 2015-2016 school year, the five third grades which got the worst math indicators for the Performance domain belong to, in order from worst to better, Phalen at Francis Scott Key 103, James Whitcomb Riley School 43, Indiana College Preparatory School, IN Math & Science Academy – South, and Riverside School 44. The first, second, and fifth school are all members of the Indianapolis Public Schools corporation. Schools in the Diocese of Gary, the Diocese of Evansville, independent, non-public schools, the Archdiocese of Indianapolis, Clay Community Schools, and South Montgomery Community Schools house the ten schools which got perfect 100%'s in the math indicator portion of the Performance domain.

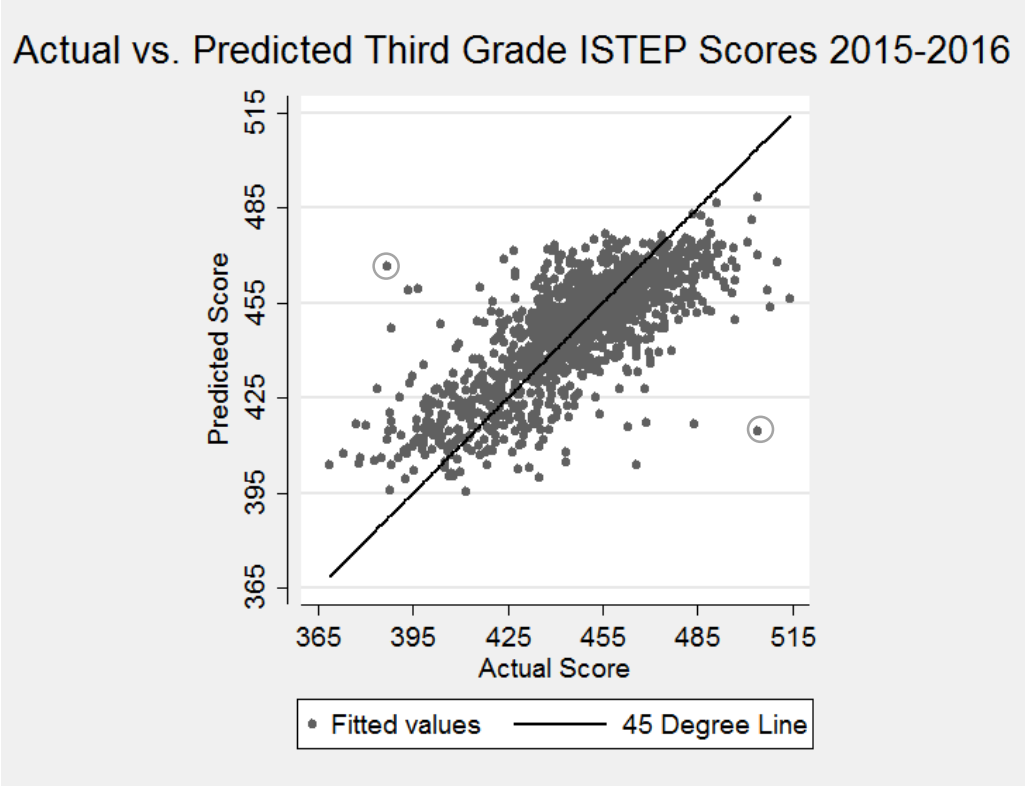
The five third grades which received the worst ELA indicators for the Performance domain for the 2015-2016 school year belong to, in order from worst to better, Indiana Christian Academy, IN Math & Science Academy – South, Jefferson Elementary School, Timothy L. Johnson Academy, and James Whitcomb Riley School 43. Note that two of these schools are the same schools which performed the worst in math as well. Two independent, non-public schools, two

[5] For the full table, see Appendix C.

Archdiocese of Indianapolis schools, and one each of Diocese of Lafayette Catholic Schools, Franklin Community School Corp, and Gary Community School Corp received Performance Domain scores of 100% for their third grades. Three schools' third grades received perfect scores in both math and ELA.

These results are interesting because it is not necessarily the poorest school districts who are performing worst for the third grade—even private schools' third grades are performing very poorly. In a way, this is good. It empirically demonstrates that there can be something about the school which influences student learning; performance is not purely demographically based.

Figure 9: Graph of Average Weighted Third Grade ISTEP Scores

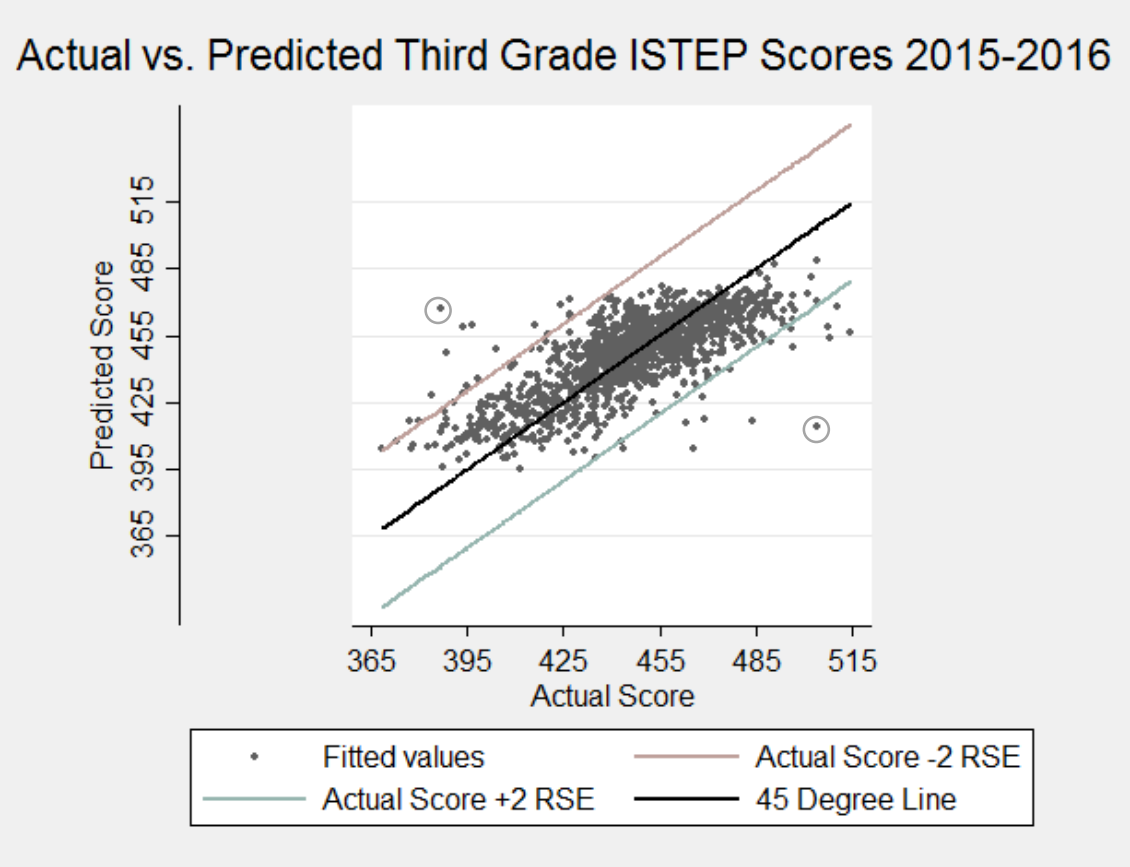


Because some of the factors which contribute to student learning and ability performance are controlled for, credit can be more accurately given to the schools for student test score outcomes. Theoretically, if a school's third grade is perfectly on par with its predicted output, its

output should fall on the black 45-degree line which runs through the graph. Falling on that line means that the school's third grade performs as expected given their demographic makeup, therefore implying that the IDOE accountability grade assigned as a result of the ISTEP scores would be relatively appropriate. Many schools fall reasonably close to the 45-degree line. There are a fair number, however, which do not. Here, 967 observations fall within +/- 1 RSE of the predicted weighted average ISTEP score. That means that roughly 75.5% of observations in my sample are captured by what I have defined as an acceptable deviation from its predicted score. The remaining 24.5%, however, deviates further from expected score.

In Figure 9 above, the further away from the line, the more the school's third grade is either under-or over-achieving. If the observation falls to the left of the line, then its actual score is worse than its predicted score. Because the predicted score accounts for demographics which contribute to the academic success or failure of students, falling left of the line implies more accurately that a school is not properly servicing its third grade. On the other hand, observations to the right of the line show that a school's actual third grade weighted average ISTEP scores are higher than its predicted ones. This indicates that a school is educating its third graders better than it was predicted to given the student demographics.

Figure 10: Accountability Designation Bands



This leads me to an alternative accountability grading system that uses the difference between an observation's actual and predicted score. Through the use of a band formed by two lines—the red and green lines in Figure 10 above—I transform the IDOE's A to F grading scale into one with three designations[6]; the band is depicted theoretically by the dashed lines in Figure 5 from Section III part F, and in practice in Figure 10 above. This band includes about 97% of the third grades in my sample. The upper and lower bars are based on deviations from the predicted ISTEP score average for this sample and are themselves calculated by taking the actual score for

[6] I do not give letter grades like the IDOE because of the tremendous variation in the sample, which is further addressed in subsequent paragraphs. I can, however, identify extreme performers with at least some level of statistical rigor. This allows me to characterize their performance based on expectations. Any degree of statistical rigor is more useful than a completely unadjusted assessment.

a given predicted score and adding +/- 2 RSE's of the mean predicted score of this sample. This is an arbitrary construction, and therefore definition, of what is "good" and "bad." I chose to create such a large band because the variation in this sample is so huge and likely to change from year to year. This way, I avoid calling truly good schools bad because they fell to the left of the forty-five-degree line in this particular sample of 2015-2016 scores.

"Exceeds Expectations" describes schools with third grade ISTEP scores to the right of the green band in Figure 9. These schools are *exceptionally* good; they are outperforming their expectations by a considerable amount and should be rewarded by the IDOE. Scoring to the right of the green band means that scores are two RSE's greater than the average predicted score for the 2015-2016 school year.

Schools to the left of red line are designated by "Does Not Meet Target." To meet this threshold, a school's third grade weighted average ISTEP scores would have to have been less than two RSE's below the average predicted score for the 2015-2016 school year. These schools are *exceptionally* bad; they underperform their expectations even after holding included influential factors constant. The IDOE needs to specially evaluate these schools.

"Exceeds Target" and "Does Not Meet Target" schools contain the top and bottom 2.5% of third grades in Indiana if the sample is normally distributed, given that the upper and lower bars are based on the addition or subtraction of two robust standard errors of the mean predicted score. In this sample, however, the number of "Exceeds Target" and "Does Not Meet Target" schools for 2015-2016 amount to 1.33% each, or 2.66% of all schools in this sample[7]. The proportion of schools which received the IDOE's top designation of an A in this sample is 20.16%. The

[7] It should be noted that this is a *random* proportion. The number of schools in "Exceeds Target" and "Does Not Meet Target" do not have to be equal, and performed on a different sample would likely be different.

proportion of schools receiving an F for this sample is 4.45%. There is a clear disparity in the IDOE's perception of school performance and my own, especially in who is considered a "top performer."

The schools within the band are considered "On Target," and essentially perform as expected. The band's purpose is to identify the exceptions to the predicted ability of the schools to turn out certain ISTEP scores which are meant to evaluate a student's acquired skills. This grading paradigm shifts the focus of improvement from an arbitrary grade to a numerical, concrete difference in actual and predicted score. As the models get more accurate through the addition of more influential variables, this grading system will become more powerful and effective.

Receiving an A designation indicates that a school is meeting all of the goals the IDOE set for it, including the progressive closure of subgroup achievement gaps and overall student growth from the previous year and performance in the current year for EMSs. When a school receives my top designation, on the other hand, it is going above and beyond its expectations. Its expectations in my model, however, are not focused on discrete gaps. I have shifted the focus of school improvement from closing *generalized* achievement gaps to maximizing student performance with the student population given to an *individual* school. Conversely, receiving an F designation tells the IDOE that a school is no where near the broad goals it needs to be meeting. My "Does Not Meet Target" designation tells the IDOE that *specific* schools are well below their potential student performance even after accounting for its student population. The distinctions are slight, but vital.

The observation circled on the right side of Figure 9 and 10 is the third grade of Benjamin Banneker Achievement Center in the Gary Community School Corporation. Its predicted third grade weighted average ISTEP score for 2015-2016 is 414.40, a score which is less than passing. In reality, it achieved a score of 504.12. This actual score puts it as the fifth highest *avg_score_w*

of all 1,280 third grades in this data set. Moreover, it performed above and beyond how it were predicted to perform. Benjamin Banneker Achievement Center should be rewarded for this achievement, because its third grade outperformed my prediction for it after controlling for race, gender, proxies for socioeconomic status, and English-language learner status.

Indiana Christian Academy is home to the third grade observation circled on the left side of the graph. This independent, non-public school's third grade has an issue opposite that of Benjamin Banneker Achievement Center's third grade. I predicted, holding demographic variables constant, that its third grade would achieve an average weighted ISTEP score of 466.66 in 2015-2016. Instead, it achieved only an *avg_score_w* of 386.69, a difference of almost 80 points. Holding race, gender, free and reduced price lunch status of students, and whether or not a student is an English-language learner constant, Indiana Christian Academy's third grade is underachieving on ISTEP. This is a prime example of how the current system masks under-performers. Figure 11 below depicts all of the exceptionally performing schools in this sample.

Figure 11: Accountability Model Comparison of Extremes

A. "Does Not Meet Target" Designated Schools

School Name	IDOE Grade	Score Differential
William McKinley School 39	F	-35.33
Jefferson Elementary School	F	-35.35
Mary Beck Elementary School	D	-36.02
Mays Community Academy	No Grade	-36.24
Adams Elementary School	B	-37.03
Marquette Montessori Academy	F	-37.12
Saint John Lutheran School		-37.40
James Whitcomb Riley School 43	F	-39.18
Community Montessori	C	-43.58
Horizon Christian Academy		-43.99
Central Christian School		-44.47
Indiana School For The Deaf		-44.53
Geist Montessori Academy	B	-44.64
Indiana College Preparatory School	D	-58.62
Hoosier Acad Virtual Charter Sch	F	-62.47
Emma Donnan Elementary School	D	-65.23
Indiana Christian Academy		-79.97

B. "Exceeds Target" Designated Schools

School Name	IDOE Grade	Score Differential
Tindley Genesis Academy	No Grade	35.30
Christel House Academy West	No Grade	35.76
Edgelea Elementary School	A	36.97
Saint Lawrence School		37.10
White Lick Elementary School	A	37.99
Indpls Lighthouse Charter School	C	38.92
Saint Paul Lutheran School		40.92
St John Paul II Catholic School		42.04
Staunton Elementary School	A	47.01
Paramount Brookside	A	47.28
Unionville Elementary School	A	48.08
Hosford Park New Tech Elementary	A	51.77
Merle Sidener Gifted Academy	A	54.32
Holland Elementary School	C	57.95
TP Schools		61.60
Frankie W McCullough Acad for Girl	A	67.42
Benjamin Banneker Achievement Ctr	C	89.72

It should be concerning that three schools received relatively good grades, but are actually underperforming so poorly that they fall in the bottom 1.33% of worst performers compared to expected performance. This further shows the importance of evaluating not only at a grade-level, but after controlling for external test score influencers. Additionally, three schools received a C from the IDOE but whose third grade performed in the top 1.33% of all schools in this sample given certain population characteristics of the school's third grade. The two tables in Figure 11 above highlight the need for controlling test score drivers beyond a school's control.

Figure 12: Inter-School Comparison Strips

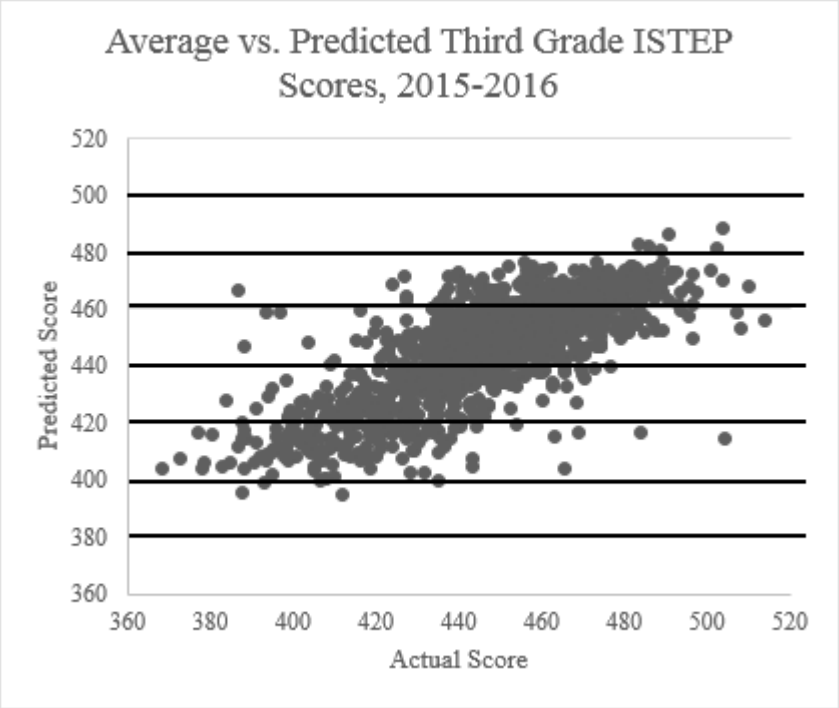


Figure 12 above represents the value of this new grading system not only to the IDOE, who can now more fairly identify the most troublesome and high-achieving schools, but to the schools themselves. The five designations of grades used in the current scoring system gave some depth to a school's performance by allowing a school to know whether it needs to improve a lot or only a little. The three designations, emphasizing once more that "On Target" encompasses roughly

97% of third grades, makes this self-reflection more difficult. To compensate for that, I have added horizontal strips as seen in Figure 12 to represent true, inter-school comparison groups.

The observations contained in each strip are predicted to do similarly based on their demographic constraints, but have an array of actual outcomes. By identifying its comparison group for a given year, a school can see how well it actually did in comparison to schools who were predicted to perform similarly. This allows schools to see how to improve themselves by analyzing same-group schools' policies, and to mimic or avoid those utilized by over- or under-achieving schools. This is much more effective than the raw, unadjusted comparisons A-F grades used IDOE currently, and allows schools to more realistically compare themselves. The IDOE uses actual scores as comparison groups, which is misleading. In the strip between 400 and 420, it is clear that Benjamin Banneker Achievement Ctr is the top performer. The other schools in this strip can see that they were predicted to perform similarly given its student population, and attempt to analyze how Banneker generated its scores. Moreover, Figure 8 above confirms that the A-F assignments are also misleading when the grade level performance is specified and after adjusting for population characteristics.

It should be noted that my model does not explain 100% of the average weighted ISTEP score, so it cannot be said that the school deserves 100% of the credit for the students' outcomes. My long model in Figure 7 explains just shy of 60% of the predicted average weighted third grade ISTEP score. Since only about 43% of the score remains unexplained and no school qualities were included in my regressions, it would make more sense for the IDOE to attribute the deviation between actual scores and predicted scores to the institutions themselves. Until we can know exactly what percentage of scores are explained by school qualities, however, we cannot assign accountability grades with full confidence. My models help increase confidence in assigning

accountability grades if 1) the state assigns grades on a grade-level basis and 2) the grade is assigned based on the difference between predicted and actual scores. At the very least, my models serve the purpose of illustrating more truthfully the role that demographics can play in ISTEP achievement (as shown in Figure 7) and how well schools are dealing with the consequences of these demographics (Figure 9).

V. CONCLUSION

In an ideal world, we know the true, exact potential of each student. We can plop an EEG on his head and see exactly what he is capable of, how much he can grow, and how likely he is to take advantage of his opportunities. Knowing this would allow us to precisely determine how schools are able to increase a student's potential and how well it accomplishes this task. We do not want to penalize schools saddled with a population of weak students and reward those schools fortunate to have a population of smart, advantaged children. We do not have a reliable quantification for individual potential, but this does not mean we cannot evaluate schools more accurately. By simply including variables which impact student performance, we can get closer to assessing a school's true performance. Controlling for anything affecting students' ISTEP scores is better than raw scores which yield a raw, unadjusted accountability grade. I hope that this novel approach to accountability grading can impact state governments nationwide, and possibly even the federal government. Introducing this sort of statistical control finally begins to address the desire for empirical evidence of improvement among students generally and student subgroups.

On a fundamental level there exist better and worse teachers, better and worse administrators, and therefore better and worse schools. My grading system is better than no control at all, but is not meant to be read in isolation. It aims, at its core, to point out that just because there

is a difference between schools' performances does not mean that it is due to the school itself. In the current system, schools, administrators, and teachers take the blame and real social issues which hinder or promote a child's learning are not addressed.

I include variables for race, English-language learning status, and socioeconomic status to get a predicted weighted average ISTEP score for third graders in Indiana. Roughly 75.5% of third grades in Indiana fall within one standard deviation above or below their predicted score for the 2015-2016 school year. The remaining 25.4% of third grades deviate more than this, which is somewhat suspicious. The difference between the predicted average score and the actual average score can give us a more focused estimation of the true influence schools have on student performance.

In my alternative way to evaluate schools, those which have actual ISTEP scores to the right of or below the band for their third grade are "Exceeds Target" target schools, indicating that the school has surpassed it's expected performance by a great deal; a school with this designation would be an *exceptionally* good one. This definition encompasses 1.33% of schools in my sample, but is subject to change between other samples. Schools within the band are "On Target," meaning that they are performing as expected given certain demographic characteristics. A school which is "On Target" year over year for its third grade is ensuring that its third grades have been performing, and by proxy learning, adequately. The schools that the IDOE really wants to scrutinize are the *exceptionally* poor ones to the left of or above the band, ones with a "Does Not Meet Target" designation. Having controlled for at least part of the factors contributing to test performance for which the school is not responsible, schools with this designation fail to perform. They could do more to improve the education of its third graders given the included independent variables. Another 1.33% of my sample's schools fall in this category and are also subject to variation

between samples. Instead of benchmarking school performance to relatively arbitrary student performance in the third grade, this different way of evaluating schools benchmarks a school to its *personalized*, baseline expected score.

To be clear, I am not proposing an entire school's accountability score be based upon only its third grade. In fact, it would be more natural to evaluate schools at the grade level. This is an additional concept important to developing a robust accountability grading system that I touch on in Section IV as well. To understand the importance of grade-level evaluations, look no further than Benjamin Banneker Achievement Center in the Gary School Corporation. In the 2015-2106 school year, its third grade had a 95.9% average passing rate for both sections of the ISTEP test. The entire school's average passing rate for both sections of the ISTEP test was only 48.6%. This difference between the school average and the grade three average is alarming; something good is happening in the third grade, perhaps within or without the school's control, and something bad is happening on a school-wide level. Assigning statistically controlled accountability scores at grade levels could more directly hold teachers and resources for that grade responsible. In future research I would like to pursue this system.

Another interesting extension of this study would be to create regressions like Figure 7 and graphs like Figure 8 for each state in the U.S. This could help give all states, as well as the federal government, an idea of how schools are performing without confounding from student population characteristics. If this were to be done, perhaps the federal government would be able to more accurately adjust their policies and incentives for social change in relation to education.

It would also be ideal to quantify and cleanly include more influences on student performance described in education and psychological literature. This would further distill the model into the effect schools have on student progress and learning. Because of random error and

omitted variable bias, it is impossible to know the exact effect of school environment by process of elimination unlike to what I allude. My grading system is not perfect because I do not know the "true" base capability level of the kids, but even a slight improvement is preferable to continuing the dysfunctional norm.

Obtaining panel data for multiple years and at a student-level broken down by classrooms and across all grades would lead to an even more empirically rigorous grading system. In conjunction with more explanatory variables, this type of data could allow for very powerful, micro-level analysis. The IDOE could then compare teachers within schools, grades between schools, and general performance between schools. This would allow a more nuanced picture of student performance to take shape, and allow for a more in-depth analysis of which components of a school are perhaps not functioning efficiently.

One drawback to this system, however, is that schools who are predicted to do poorly and are "On Target" are not pointed out, yet they have students who are performing poorly on ISTEP. If policymakers want to improve these schools in the bottom left corner—the ones "On Target" with predictions of low scores—they need to fix the root causes of the prediction and subsequent performance. These causes are partially captured in the explanatory variables of my regression, and discussed more fully in Section II. The current system is a hack at improving student outcomes, but largely ignores many root causes of student outcomes.

As Calabresi and Melamed (1972) put it, my analysis is only "one view of the cathedral." They remind readers that Monet painted a portrait of the cathedral at Rouen many times, under different light and conditions, because it cannot be captured it in one take. So, too, there are many facets of school accountability; my method only begins to scratch the surface of the expanse that is maintaining and improving student educational outcomes.

It is a lens through which to view this byzantine system of accountability grading both in the United States as a whole and in Indiana specifically. I offer here a view which allows educators, policymakers, and public citizens to take one step closer towards giving due credit to schools in Indiana's third grade standardized test results. My hope is that this one step will reveal a path leading to the creation of a truly rigorous, objective accountability system for schools in Indiana and across the nation. The emphasis on accountability for student outcomes creates dangerous incentives for "struggling" school systems. If you do not believe me, reminisce on the Atlanta Public Schools scandal of 2009 and 2010, or think about the fact that Texan educators consistently teach to their state's standardized test to avoid intervention.

President Bush rightly scorned "the soft bigotry of low expectations." We should want to employ these high-stakes testing requirements and aggressive accountability standards, but only if we can appropriately judge the responsibility of our schools. What better, more objective way is there to do so than to exploit modern statistical techniques and the data we have at our fingertips?

VI. APPENDICES

Appendix A: Entire Fifty State Accountability Scoring Analysis

Please follow this link to see the complete analysis:

https://emilytroyer2019.wixsite.com/mysite/post/appendix-a-entire-fifty-state-accountability-scoring-analysis.

Appendix B: Income Eligibility Guidelines for 2015-2016 from the IDOE

INCOME ELIGIBILITY GUIDELINES*
EFFECTIVE FROM JULY 1, 2015 – JUNE 30, 2016

	Reduced Price Meals 185% of federal poverty guidelines					Free Meals 130% of federal poverty guidelines				
Household Size	Yearly	Monthly	Twice Per Month	Every Two Weeks	Weekly	Yearly	Monthly	Twice Per Month	Every Two Weeks	Weekly
1……	21,775	1,815	908	838	419	15,301	1,276	638	589	295
2……	29,471	2,456	1,228	1,134	567	20,709	1,726	863	797	399
3……	37,167	3,098	1,549	1,430	715	26,117	2,177	1,089	1,005	503
4……	44,863	3,739	1,870	1,726	863	31,525	2,628	1,314	1,213	607
5……	52,559	4,380	2,190	2,022	1,011	36,933	3,078	1,539	1,421	711
6……	60,255	5,022	2,511	2,318	1,159	42,341	3,529	1,765	1,629	815
7……	67,951	5,663	2,832	2,614	1,307	47,749	3,980	1,990	1,837	919
8……	75,647	6,304	3,152	2,910	1,455	53,157	4,430	2,215	2,045	1,023
For each additional person:	+7,696	+642	+321	+296	+148	+5,408	+451	+226	+208	+104

*For the 48 contiguous United States, District of Columbia, Guam and territories

FOR SCHOOL USE ONLY
– NOT TO BE DISTRIBUTED TO HOUSEHOLDS

Appendix C.: Major Accountability Grading Discrepancies—School-Level vs. Grade-Level

Avg Weighted Third Grade ISTEP Score	IDOE Grade	My Grade
472.9615479	A	D
460.7703552	A	D
453.7352905	A	D
460.0659485	A	D
466.6457214	A	D
462.2228699	A	D
457.4383545	A	D
460.5	A	D

460.3721008	A	D
457.6499939	B	F
456.4744263	A	D
450.125	A	D
457.527771	A	D
447.5333252	B	F
450.0131531	A	D
450.0322571	B	F
451.9264832	A	D
448.359375	A	D
452.0119019	A	D
452.8373413	A	D
440.4615479	B	F
442.0208435	B	F
447.3657532	A	D
442.5138855	B	F
442.9549561	A	D
439.7580566	A	D
444.2926941	A	D
450.3529358	B	F
445.4552307	A	D
447.0505066	A	D
451.0654907	A	D
444.0491943	A	D
438.3265381	B	F
439.7037048	B	F
448.372345	A	D
441.0606079	B	F
448.4534302	A	D
442.7538452	B	F
443.5115051	B	F
436.4341431	B	F
447.07547	B	F
445.2826233	B	F
445.6190491	B	F
442.3500061	A	D
448.3392944	A	D
445.8863525	A	D
439.0782166	B	F
439.2999878	A	D
443.5138855	B	F
443.1739197	A	D
441.0454407	B	F
439.6142883	B	F
441.7592468	B	F
442.6203308	B	F
445.0112305	A	D
437.3796387	B	F

441.7814636	B	F
441.9024353	B	F
443.9240417	A	D
430.2536316	B	F
443.6363525	A	D
440.4886475	A	D
436.2391357	B	F
437.8399963	A	D
439.5576782	B	F
444.2434692	A	D
437.6639404	B	F
438.8253479	B	F
444.6000061	A	D
437.9304199	B	F
437.946228	B	F
433.296875	B	F
435.1634521	B	F
437.5649414	B	F
442.8023376	A	D
435.14151	B	F
434.980011	B	F
435.4657593	B	F
438.6395264	B	F
438.2633972	B	F
438.4351196	B	F
431.4076233	B	F
438.3970642	B	F
438.9714355	A	D
428.1711121	B	F
438.1931152	A	D
440.4685364	B	F
432.7182312	B	F
439.7600098	B	F
435.5109558	B	F
428.6764832	B	F
433.4095154	B	F
435.2888794	B	F
435.1397705	B	F
443.9468079	A	D
434.5100098	B	F
435.25	B	F
435.4958801	B	F
436.6976624	A	D
441.2484741	B	F
434.2597961	B	F
441.0593262	B	F
436.253418	B	F
429.664917	B	F

432.5067444	B	F
434.9874878	B	F
432.5329895	B	F
436.2045593	B	F
436.2512207	B	F
432.2171631	B	F
434.1080933	B	F
430.4553223	B	F
435.175293	B	F
431.2272644	B	F
434.4423218	B	F
432.0961609	B	F
429.6303101	B	F
433.1696472	B	F
429.4727173	B	F
430.2250061	B	F
439.1881104	A	D
436.3936157	A	D
434.9360352	B	F
432.008606	B	F
437.123291	B	F
423.9818115	B	F
430.2428589	B	F
434.4961243	B	F
427.2377014	B	F
425.2763062	B	F
429.3609924	B	F
423.9040527	B	F
431.6714172	B	F
423.8118896	B	F
426.3333435	B	F
430.6818237	B	F
434.8863525	B	F
419.0278931	B	F
422.9861145	B	F
421.4046326	B	F
428.5198364	B	F
423.3181763	B	F
420.8812561	B	F
427.7012939	B	F
435.1323547	A	D
423.8800049	B	F
429.4513855	B	F
425.884613	B	F
419.9857178	B	F
427.0530396	B	F
415.6511536	B	F

420.6756897	B	F
418.8959351	B	F
424.0119019	B	F
427	B	F
422.0714417	B	F
425.2970886	B	F
421.7642822	B	F
415.7290344	B	F
417.0827637	B	F
404.375	B	F
411.4562378	B	F
417.3725586	B	F
409.6190491	B	F
409.7799988	B	F
409.1135254	B	F
398.2666626	B	F
398.2561035	B	F
447.2528076	A	F
443.1363525	A	F
439.8921509	A	F
445.4338379	A	F
441.0816345	A	F
441.7058716	A	F
447.1666565	A	F
433.625	A	F
432.8048706	A	F
442.1694946	A	F
433.7322693	A	F
436.307251	A	F
442.4230652	A	F
440.1826782	A	F
428.3945007	A	F
433.8269348	A	F
427.5935974	A	F
423.3218384	A	F
416.1080933	A	F
418.8676453	A	F
428.6790161	A	F
420.2990723	A	F
421.2580566	A	F
419.5294189	A	F
415.6521606	A	F

Appendix D. Gender Influences on Math and ELA ISTEP Scores 2015-2016

Gender Influence on ISTEP Scores

	Math Component			ELA Component		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
males	-8.3195	-10.6292	-7.2769	-30.0472***	-32.0862***	-28.7427***
	(9.6e+00)	(8.6e+00)	(8.1e+00)	(7.7e+00)	(6.9e+00)	(6.3e+00)
ell		-4.6829	4.6593		-9.7386	-1.2300
		(6.8e+00)	(6.7e+00)		(5.8e+00)	(5.7e+00)
not_white		-50.5703***	-22.5874***		-38.3583***	-11.2901***
		(3.0e+00)	(3.4e+00)		(2.5e+00)	(2.8e+00)
free_lunch			-54.0652***			-51.4573***
			(3.1e+00)			(2.4e+00)
rp_lunch			-17.9228			-7.3558
			(1.2e+01)			(8.2e+00)
_cons	441.9008***	458.9181***	473.0694***	469.2908***	482.7242***	495.3277***
	(4.9e+00)	(4.4e+00)	(4.3e+00)	(3.9e+00)	(3.5e+00)	(3.3e+00)
N	1280	1280	1280	1280	1280	1280
R^2	0.001	0.319	0.490	0.012	0.318	0.555
F	0.7493	219.3507	199.9873	15.1581	227.3838	270.6556

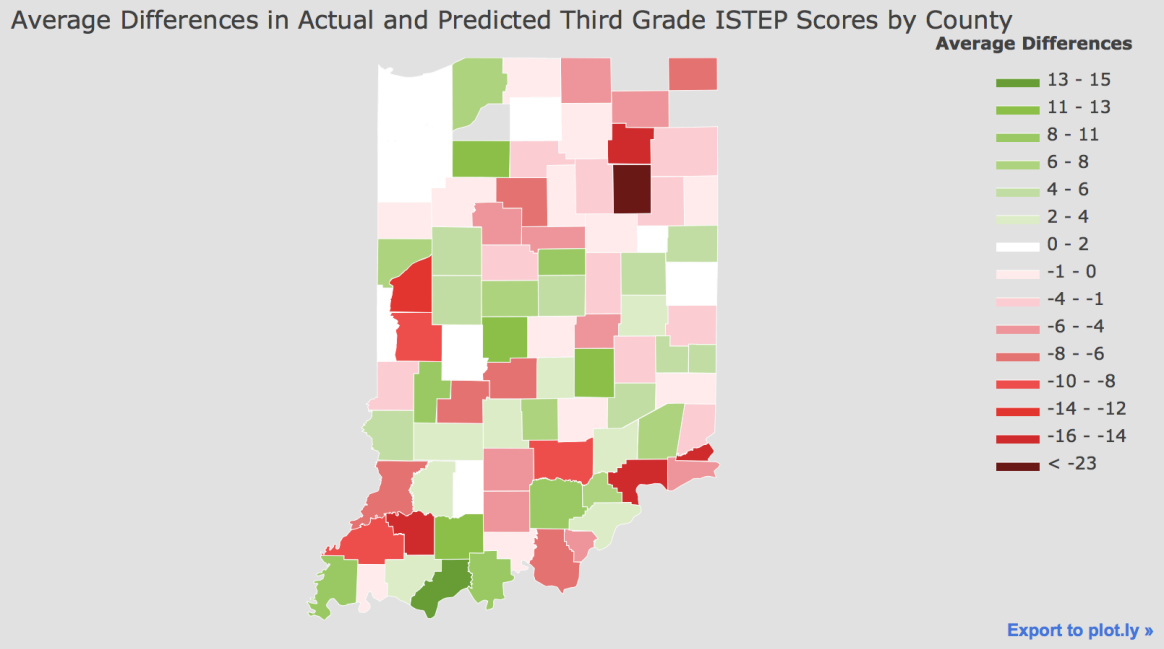
Robust standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

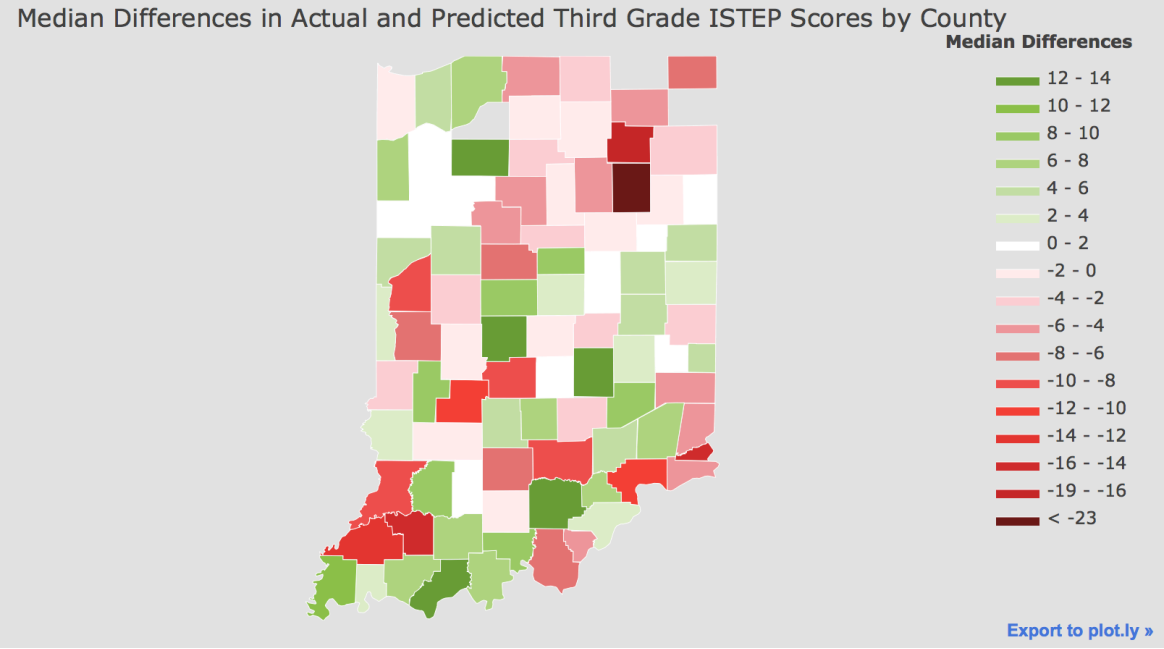
Model 2 and Model 5 leave out *free_lunch* and *rp_lunch* to show the effect of gender before and after controlling for socioeconomic status. Almost all previous literature, including Reardon, Fahle, Kalogrides, Podolsky, and Zárate (2018)'s new working paper, show that females outperform males in math. My data shows this. However, my data does not support Reardon, Fahle, Kalogrides, Podolsky, and Zárate (2018) on the math component side because even when I control for socioeconomic status, the effect of being male on ISTEP's math portion is not significant.

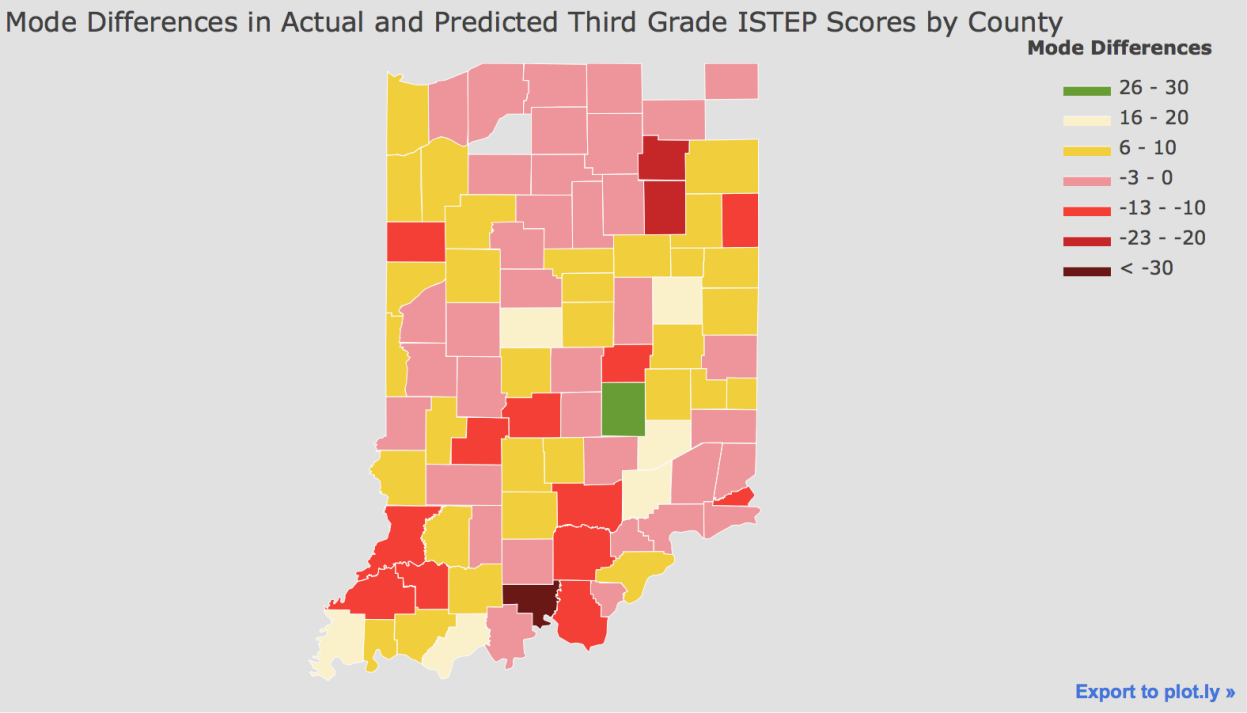
Appendix E: Using Python to Model Third Grade Differentials—Indiana Maps by County

Follow this link for the live blogpost which is a more informal, shortened commentary on the subject matter of this thesis and contains live maps of the average, median, and mode difference between third grades' actual and predicted scores by county. The link to the blog post is here:

https://emilytroyer2019.wixsite.com/mysite/post/improving-indiana-s-accountability-scoring-system. The maps are interactive and report various data points when you hover over a county. They were constructed in Python's Jupyter Notebooks using primarily plot.ly and Pandas libraries. Stills of these maps are shown below.







Appendix F: School Designations in the New System

Please see this link for the full listing: https://emilytroyer2019.wixsite.com/mysite/post/school-designations-in-the-new-system.

VII. BIBLIOGRAPHY

"50 States Comparison." (2018). *Education Commission of the States.* Retrieved from http://ecs.force.com/mbdata/mbQuest6S?rep=SA171

Abernathy, S. F. (2008). *No child left behind and the public schools*. Retrieved from https://ebookcentral.proquest.com

"Accountability." (n.d). *Indiana Department of Education*. Retrieved from https://www.doe.in.gov/sites/default/files/essa/accountabilitry-summary.pdf

Bæk, U. (2016) Rural Location and Academic Success—Remarks on Research, Contextualisation, and Methodology. *Scandinavian Journal of Educational Research, 60*(4), 435-448.

Barton, P. E. and Coley, R. J. (2009). Parsing the Achievement Gap II. *Educational Testing Services—Policy Evaluation and Research Center*. Retrieved from https://www.ets.org/Media/Research/pdf/PICPARSINGII.pdf

Brenchley, C. (2015, April 8). What is ESEA? *Homeroom*. Retrieved from https://blog.ed.gov/2015/04/what-is-esea/

Caldas, S., & Bankston, C. (2005). *Forced to fail: the paradox of school desegregation*. Retrieved from https://ebookcentral.proquest.com

"Census estimates show Indiana growing more diverse, older." (2014, June 26). *IU Bloomington News Room*. Retrieved from http://archive.news.indiana.edu/releases/iu/2014/06/census-estimates-race-ethnicity.shtml

Chatterji, M. (2006). Reading Achievement Gaps, Correlates, and Moderators of Early Reading Achievement: Evidence From the Early Childhood Longitudinal Study (ECLS)

Kindergarten to First Grade Sample. *Journal of Educational Psychology, 98*(3), 489–507. Retrieved from https://doi.org/10.1037/0022- 0663.98.3.489

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have Gender Gaps in Math Closed? Achievement, Teacher Perceptions, and Learning Behaviors Across Two ECLS-K Cohorts. *AERA Open, 2*(4), 233285841667361. Retrieved from https://doi.org/10.1177/2332858416673617

Consequences, Ind. Code 20-3, Chapter 9 Section b-c. (2018).

Education Commission of the States. (2018, May). 50 State Comparison. Retrieved from http://ecs.force.com/mbdata/mbQuest6S?rep=SA171

Education Commission of the States. (2018, May). History. Retrieved from https://www.ecs.org/about-us/history/

Elementary and Secondary Education Act of 1965. Pub. L. 89-10, 79 Stat. 79. Retrieved from https://www.scribd.com/doc/49149656/Elementary-and-Secondary-Education-Act-of-1965#download&from_embed); https://www2.ed.gov/policy/elsec/leg/esea02/pg2.html #sec1111

Epstein, J. (2011). School, family, and community partnerships: Preparing educators and improving schools. Boulder, CO: Westview Press.

"ESSA Amendment Accountability Summary." (n.d). *Indiana Department of Education*. Retrieved from https://www.doe.in.gov/sites/default/files/essa/essa-amendment-overview.pdf

"Every Student Succeeds Act (ESSA)." (2018). *Maine Department of Education.* Retrieved from https://www.maine.gov/doe/Testing_Accountability/ESSA

"Every Student Succeeds Act (ESSA)." (n.d.) *U.S. Department of Education.* Retrieved from

https://www.ed.gov/essa

Fennema, E. & Sherman, J. (1977). Sex-Related Differences in Mathematics Achievement, Spatial Visualization and Affective Factors. *American Educational Research Journal, 14*(1), 51-71.

Frisby, C. L. (2013). *Meeting the psychoeducational needs of minority students: evidence-based guidelines for school psychologists and other school personnel*. Retrieved from https://ebookcentral.proquest.com

Fryer, R. G., & Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics, 2*(2), 210–240. Retrieved from https://doi.org/10.3386/w15430

Gewerts, C. (2015, December 3). No Child Left Behind Rewrite Spells Out Assessment Details. *Education Week.* Retrieved from http://blogs.edweek.org/edweek/curriculum/2015/12/no_child_left_behind_rewrite_spells_out_assessment_details.html

Grew, J. R. & Sheldrake, W. J. (Presenters). (2013, September 6). Examination of Indiana's A to F School Accountability Model. *Indiana Speaker of the House Brian Bosma and Indiana Senate President Pro-Tempore David Long*. Retrieved from http://www.in.gov/legislative/pdf/Accountability_Model_Exam.pdf

Grodsky E., Warren, J.R., & Felts, E. (2008). Testing and Social Stratification in American Education. *Annual Review of Sociology, 34*, 385-404.

Hanushek, E. & Peterson, P. (2019, March 17). The War on Poverty Remains a Stalemate. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/the-war-on-poverty-remains-a-stalemate-11552847932.

Haskins, A. R. & Jacobsen, W. C. (2017) Schools as Surveilling Institutions? Paternal

Incarceration, System Avoidance, and Parental Involvement in Schooling. *American Sociological Review, 82*(4), 657-684.

Helms, J.E. (1992, September). Why is There No Study of Cultural Equivalence in Standardized Cognitive Ability Testing? *American Psychologist, 47*(9), 1083-1101.

Hermes, P. (2019, March 26). Time to Address the Real Poverty-Gap Issues. *Wall Street Journal.* Retrieved from https://www.wsj.com/articles/time-to-address-the-real-poverty-gap-issues-11553632502

Hernandez, D. J. (2011, April). Double Jeopardy: How Third Grade Reading Skills and Poverty Influence High School Graduation. *The Annie E. Casey Foundation*.

"History." (2018). *Education Commission of the States.* Retrieved from https://www.ecs.org /about-us/history/

Hood, J. (2011). CPS fails to close performance gap. *The Chicago Tribune.* Retrieved from https://www.chicagotribune.com/news/ct-xpm-2011-11-14-ct-met-cps-racial-gap-1114-20111114-story.html

"How Are the Local, State And Federal Governments Involved In Education? Is This Involvement Just?" (2019). *The Center for Public Justice*. Retrieved from https://www.cpjustice.org/public/page/content/cie_faq_levels_of_government

Husain, M., & Millimet, D. L. (2009). The mythical "boy crisis"? *Economics of Education Review, 28*(1), 38– 48. Retrieved from https://doi.org/10.1016/j.econedurev.2007.11.002

Indiana Department of Education. (2017). ISTEP+ Indiana Statewide Testing for Educational Progress—Plus: 2017 Guide to Test Interpretation. *Indiana Department of Education*. Retrieved from https://www.doe.in.gov/sites/default/files/assessment/2017-istep-guide-test-interpretation20.pdf

Item Response Theory. (2019). Retrieved from https://www.mailman.columbia.edu/research/population-health-methods/item-response-theory

Jeffrey, J. (1978). Education for children of the poor: A study of the origins and implementation of the Elementary and Secondary Education Act of 1965. Columbus: Ohio State University Press

Jencks, C., & Phillips, M. (Eds.). (2006). *The black-white test score gap*. Retrieved from https://ebookcentral.proquest.com

Lindsay, Jeanie. (2019). What does it mean for a school to 'fail'? *Indiana Public Broadcasting Stations.* Retrieved from https://indianapublicmedia.org/stateimpact/tag/takeover-schools/

McCormick, J. (n.d.). Indiana Student Centered A-F Accountability System. *Indiana Department of Education*. Retrieved from https://www.doe.in.gov/sites/default/files/accountability/f-accountability-presentation.pdf

Mosle, S. (2018, September 6). Raising Student Performance the Right Way: Can Good Teaching Be Taught? *The New York Times Magazine*. Retrieved from https://www.nytimes.com/interactive/2018/09/06/magazine/student-performance-atlanta-teaching.html

"Overview of the State Accountability Report Card." (2018, August 24). *California Department of Education*. Retrieved from https://www.cde.ca.gov/ta/ac/sc/overview.asp

Paul, C. A. (2016). Elementary and Secondary Education Act of 1965. *Social Welfare History Project*. Retrieved from http://socialwelfare.library.vcu.edu/programs/education/elementary-and-secondary-education-act-of-1965/

Primi et. al. (2016). The Development and Testing of a New Version of the Cognitive Reflection

Test Applying Item Response Theory (IRT). *Journal of Behavioral Decision Making, 29*, 453-469.

Reardon, S.F., Fahle, E.M., Kalogrides, D., Podolsky, A., & Zárate, R.C. (2018). Gender Achievement Gaps in U.S. School Districts (CEPA Working Paper No.18-13). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp18-13

Ritz, G. (2017). English Learner Guidebook, 2016-2017. *Indiana Department of Education, Office of English Learning & Migrant Education*. Retrieved from http://www.doe.in.gov /sites/default/files/elme/2016-2017-el-guidebook.pdf

Ruszkowski, C. N. (2017). New Mexico A-F School Grading Technical Guide: Calculations and Business Rules for Schools and Districts. *New Mexico Public Education Department, Assessment and Accountability Division*. Retreived from http://aae.ped.state.nm.us/ SchoolGradingLinks/1617/Technical%20Assistance%20for%20Educators/Technical%20Guide%202017.pdf

"School Accreditation." (2017, April 5). *Indiana Department of Education*. Retrieved from https://www.doe.in.gov/accountability/school-accreditation

Sibley, E. & Dearing, E. (2014). FAMILY EDUCATIONAL INVOLVEMENT AND CHILD ACHIEVEMENT IN EARLY ELEMENTARY SCHOOL FOR AMERICAN-BORN AND IMMIGRANT FAMILIES. *Psychology in the Schools, 51*(8), 814-831.

"The ABCs of ESEA, ESSA and No Child Left Behind." (2019). *Education Post.* Retrieved from https://educationpost.org/the-abcs-of-esea-essa-and-no-child-left-behind

"The Big Idea of School Accountability." (2015) *George W. Bush Institute.* Retrieved from https://www.bushcenter.org/essays/bigidea/.

Valdez-Pierce, L. (2003). *Assessing English Language Learners.* Washington, D.C.: National

Education Association of the United States.

Weddle, E. (Producer). (2019, February 28). *The Takeover* [Audio podcast]. Retrieved from https://www.wfyi.org/news/articles/thetakeover