STUDENT LEARNING, STUDENT DEMOGRAPHICS, OR SOMETHING ELSE?
A QUANTCRIT ANALYSIS OF HOW SCHOOL ACCOUNTABILITY REFLECTS
STUDENT LABELS BUT NOT STUDENT NEEDS

by
Kim Strong
B.A., Portland State University, 2009
M.A.T., North Carolina State University, 2013

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
School of Education
2022

Committee Members:
Kathy Escamilla
Mimi Engel
Kira Hall
Sue Hopewell
Michelle Renée Valladares
Abstract

Kim Strong, (Ph.D., School of Education)

Student Learning, Student Demographics, Or Something Else? A QuantCrit analysis of how school accountability reflects student labels but not student needs

Thesis directed by Professor Emeritus Kathy Escamilla

Improving educational outcomes and school conditions for historically marginalized students has been a primary goal since the school accountability movement began in the 1960s. However, despite decades of legislation, policy, and enactment designed to achieve this purpose, historically marginalized students continue to suffer from disparate academic outcomes. Using Critical Race Theory and QuantCrit frameworks, this dissertation analyzed accountability outcomes in relation to student demographics, school contexts, and English Learner characteristics and services in Denver Public Schools over three years to understand what the accountability framework employed by the district was and was not measuring. Findings indicate that the schools with the highest accountability ratings consistently had (a) smaller proportions of students of color, students receiving Free and Reduced Lunch services, and English Learners; (b) higher rates of Fully Qualified teachers and students identified as Gifted and Talented; and (c) nearly half the frequency of disciplinary actions, incidents, and actions resulting in instructional loss. When these variables were used in Ordinary Least Squares (OLS) and ordinal logit multiple regressions, this study revealed that student demographics and disciplinary actions were statistically significant predictors of both accountability scores and outcomes. These results indicate that the accountability framework used by the district was biased in favor of schools that
served small proportions of historically marginalized students while ignoring and hence failing to address disparate access to educational resources like high quality teachers, Gifted and Talented programs, native language supports, and less punitive disciplinary environments. These failures to measure and thus encourage equitable learning environments coincided with a downward trend of schools increasingly gaining failing accountability status during the study, with charter schools – which some see as solutions to public school dysfunctions – having the highest rates of discipline and lowest rates of language supports for English Learners. Implications of this study include the recommendation that districts conduct “equity reviews” to ensure accountability policies do not disproportionately harm historically marginalized students and that accountability frameworks include metrics to evaluate school contexts and services to promote the equitable allocation of resources and opportunities to all students.
Dedication

This work is dedicated to the many people whose support, wisdom, generosity, kindness, and guidance allowed me to conduct this work.

Me gustaría agradecer a mi familia en Carolina del Norte: María, Chofo, Jared, Isa, Flaco, Sonya, Byron, Yetnaletzi, Chayo, Maje, Ashira, Aye, Edith, Jeremy, Mitzy, Cristal, Dulce, Laila y Panchito. Desde cuando nos conocimos me hicieron sentir parte de la familia. Gracias por enseñarme lo que es tener un gran corazón y una familia como Dios manda. En particular quiero agradecerle a Mitzy, quién me inspiró a hacer el doctorado en educación. Mitzy, mereces muchísimo más que lo que has recibido de las escuelas y eso es culpa del sistema – no tuya. Con mucho gusto espero ver todas las grandes cosas que tú y tu familia logren. Este trabajo es para ustedes por compartir sus historias conmigo y dejarme ser parte de sus vidas.

To my Colorado family I owe more gratitude than I can express, not just for what you have done for me but for my family as well. You welcomed us into your home for holidays, birthdays, and barbeques. You celebrate our triumphs and share in our tears. You have been Katherine’s ‘buelos, so that neither she nor we feel so alone here. From the bottom of my heart, thank you for being so good to us. Los queremos muchísimo, don Manuel y doña Kathy. Thank you for being our family. Kathy, I could not have gotten through this program without you. You have been my idea sounding board, my encouragement to move forward during difficult times, my role model, my mentor, and my friend. If I could achieve a fraction of the brilliance, kindness, goodness, and service that is your legacy, I will consider my life a success. You are the best person I know, and the lessons I have learned from you have made me who I am today. Anything that I may accomplish will be in no small part due to the impact you have had in my life and that of my family. Thank you for being the best advisor, role model, and friend I could hope for. I humbly dedicate this work to you as a token of gratitude for refusing to give up on me.

Finally, this dissertation is dedicated to Katherine and Simitrio, to whom I owe everything. Not just for your patience through the long nights and missed weekends of graduate school, but because you have given me a reason to be here. Thank you for the (many!) times you both encouraged me to just drop out, saying I didn’t need a PhD for you to be proud of me. You have taught me what real love is. Thank you for showing me what it is to have a family, to care and be cared for, to know contentment and peace beyond what I had imagined was possible. You are my world, and I love you with all my heart.
Acknowledgments

This research was conducted in support of the decades-long advocacy of the Congress of Hispanic Educators. Thank you for allowing me to contribute to your historic fight for the educational rights of bilingual students and families. Special thanks to Dr. Martha Urioste, Esther Romero, Dr. Darlene LeDoux, Roger Rice, and Lu Liñana (in memoriam) for your mentorship and encouragement throughout the years.

Without a doubt, I could not have done this work without my committee, and to them I owe a special debt of gratitude. More than act as examples of outstanding scholarship and complex thinking, more than model professional talents and accomplishments, they exemplify how to be fundamentally good people. From them, I have learned what it means to be a true scholar: unapologetically brilliant yet not condescending, leaders of their fields who make time for mentorship, exceptionally accomplished and still kind-hearted. I have learned from each of you so much more than academic skills – you have taught me how to be a strong woman with a clear sense of purpose, both in academia and beyond, and that is a lesson that will stay with me for a lifetime. Thank you, Dr. Kathy Escamilla, Dr. Mimi Engel, Dr. Kira Hall, Dr. Sue Hopewell, and Dr. Michelle Renée Valladares.
# Table of Contents

LIST OF TABLES ................................................................................................................................. VIII

LIST OF FIGURES ................................................................................................................................. IX

INTRODUCTION .......................................................................................................................................... 1

Sociohistorical Context ........................................................................................................................... 1

Study Site Context .................................................................................................................................. 12

Theoretical Framework ........................................................................................................................... 18

Conceptual Framework ......................................................................................................................... 22

Purpose of Study ................................................................................................................................... 33

Research Questions ................................................................................................................................. 34

Significance of the Research ................................................................................................................... 35

LITERATURE REVIEW ............................................................................................................................. 37

Research Regarding the Validity of Accountability Policies ................................................................. 38

Research Regarding the Efficacy and Outcomes of Accountability Policies for Historically Marginalized Students .................................................................................................................. 43

Quantcrit Approaches To Understanding School Environments and Outcomes for Historically Marginalized Students ................................................................................................................... 47

Relationship Between Previous Research and the Dissertation ......................................................... 52

METHODS .............................................................................................................................................. 55

Data Overview and Study Parameters .................................................................................................... 55

Research Process ..................................................................................................................................... 64

Methods per Research Question (RQ) ...................................................................................................... 66

Analysis and Interpretation ..................................................................................................................... 74

Positionality Statement ............................................................................................................................ 74

RESULTS ............................................................................................................................................... 77
RESEARCH QUESTION 1: WHAT ARE THE STUDENT DEMOGRAPHICS, EL CHARACTERISTICS, AND SCHOOL CONTEXTS PER SCHOOL PERFORMANCE FRAMEWORK (SPF) RATING BRACKET? ................................................................. 77

RESEARCH QUESTION 2: AT WHAT RATE DID SCHOOLS REMAIN IN, ENTER INTO, OR EXIT THE MOST EXTREME SPF RATINGS DESIGNATIONS OF INTERVENTION AND BLUE STATUS, AND WHAT ARE THE STUDENT DEMOGRAPHICS, EL CHARACTERISTICS AND SERVICES, AND SCHOOL CONTEXTS IN THESE STATUSES? ......................................................... 85

RESEARCH QUESTION 3: WHAT ARE THE STUDENT DEMOGRAPHICS, EL CHARACTERISTICS AND SERVICES, AND SCHOOL CONTEXTS PER CHARTERS AND DISTRICT-RUN SCHOOLS? ............................................................................................................ 93

RESEARCH QUESTION 4: DO STUDENT DEMOGRAPHICS PREDICT PERCENT SPF POINTS EARNED? .......................................................... 98

RESEARCH QUESTION 5: DO STUDENT DEMOGRAPHICS PREDICT SPF OUTCOMES? .......................................................................................... 111

SUMMARY .......................................................................................................................... 116

DISCUSSION ......................................................................................................................... 117

IMPLICATIONS FOR RESEARCHERS .................................................................................. 127

IMPLICATIONS FOR TEACHERS, FAMILIES, AND ADVOCATES ........................................ 130

LIMITATIONS ..................................................................................................................... 132

CONCLUSION ....................................................................................................................... 133

REFERENCES ....................................................................................................................... 137

APPENDIX A ......................................................................................................................... 161

APPENDIX B ......................................................................................................................... 163
List of Tables

Table 1. Pearson Correlations of Potential Control Variables and SPF Percent Points Earned ................................................................. 70

Table 2. Pearson Correlation of Student Demographic Predictors and SPF Percentage Used in Multiple Regressions ............................................. 72

Table 3. Means of Student Demographics, English Learner Characteristics, Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Years 2016-2017 through 2018-19 .................................................. 78

Table 4. Descriptive Statistics of Schools that Remained In, Entered Into, and Exited From Intervention Status and Blue Status per District-Run and Charter Schools as of the Final Year of the Study (2018-2019) ................................................ 86

Table 5. Key To Abbreviated Variable Names ................................................................. 89

Table 6. Descriptive Statistics of Means of Schools that Remained In, Entered Into, and Exited From Intervention Status and Blue Status Across the Three-Year Study Timeframe Aggregate ................................................................. 90

Table 7. Means of Study Variables per District-Run And Charter Schools for Each Year of Study and Three-Year Aggregate ........................................ 94

Table 8. Individual Predictor and Saturated Models OLS Regressions with Cubed Terms for Academic Years 2016-2017 Through 2018-2019 ..................... 100

Table 9. Descriptive Statistics of Percentiles, Minimum and Maximum Values, Standard Deviations, and Means for Each Student Demographic Predictor Variable Used in Multiple Regressions ........................................ 106

Appendix Tables

Table 1. Data Sources, Datasets, Data Types, and Data Uses in Dissertation ............... 161

Table 2. Means of Student Demographics, English Learner Characteristics, English Learner Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Year 2016-2017 .................................................. 163

Table 3. Means of Student Demographics, English Learner Characteristics, English Learner Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Year 2017-2018 .................................................. 164

Table 4. Means of Student Demographics, English Learner Characteristics, English Learner Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Year 2018-2019 .................................................. 165
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.</td>
<td>Denver Public Schools SPF Color-Coded Rating Brackets Description And Points Cutoffs</td>
<td>14</td>
</tr>
<tr>
<td>Figure 2.</td>
<td>Scatterplots Panels of Student Demographics and the Percent of SPF Points Earned</td>
<td>71</td>
</tr>
<tr>
<td>Figure 3.</td>
<td>Mean Percentages of Select Student Demographics, EL Characteristics and Services, and School Contexts Across SPF Ratings Brackets for Each Year</td>
<td>79</td>
</tr>
<tr>
<td>Figure 4.</td>
<td>Predicted Percent SPF Points Earned per Individual Student Demographic Variables Reflecting Models 2, 4, 6, and 8</td>
<td>106</td>
</tr>
<tr>
<td>Figure 5.</td>
<td>Predicted Probabilities of Receiving Simplified SPF Outcomes (A) Intervention, (B) On-Watch, or (C) High Performing per Student Demographic Predictor Using Models 2, 4, 6, and 8 from Research Question 4</td>
<td>113</td>
</tr>
</tbody>
</table>
Introduction

Sociohistorical Context

Legislative History

Although assessments have been used in the United States since the nineteenth century, how they embodied “accountability” was very different from our contemporary understanding. Historically, assessments held students, rather than teachers, accountable for their own learning (Ravitch, 2002). Public oversight of education was manifested through school board elections and the input-oriented reporting of funding allocations to ensure that all students were provided with adequate resources (Cuban, 2004; Elmore & Fuhrman, 1995). This changed in the early twentieth century, as pseudo-scientific standardized testing gained popularity as a part of the eugenics movement (Zuberi, 2001) and the newly formed departments of education throughout US colleges began to see assessments as scientific instruments which could precisely measure learning and progress (Ravitch, 2002).

Although professional educators at the time adhered to a belief that the purpose of education was to prepare future citizens and that educational shortcomings would be best remediated through improved support (Ravitch, 2002), this orientation was dramatically altered with the passage of the Elementary and Secondary Education Act (ESEA) in 1965. The ESEA was remarkable, not only for being a sweeping piece of federal legislation specifically designed to improve the education of students of color and students in poverty (DeBray-Pelot & McGuinn, 2009; Thomas & Brady, 2005), but also because it represented a turning point by tying federal funding to measurable evidence of program effectiveness, thus setting the stage for testing to
become the measure of school success and the bedrock of the modern school accountability movement (Cuban, 2004; Ryan & Shepard, 2008).

This focus on outputs like test scores was further cemented with the release of the 1966 Coleman Report, a Congressionally-mandated study which focused on academic outputs like test scores rather than inputs like supports, concluding that school funding and resources alone were not predictive of academic achievement but rather a combination of other variables such as family background and school composition were more strongly correlated with outcomes (Coleman et al, 1966). Despite employing problematic methodology (Darling-Hammond, 2004), the Coleman Report had an outsized impact on the popular understanding of education reform, with its recommendation to desegregate lost to the deficit view that families in poverty and families of color were variables that correlate to student academic failure (Ladson-Billings, 2006), a failure which no amount of additional funds or resources could ameliorate (Hanushek, 1997).

The Coleman Report also ushered in a new public attention to educational achievement, and by the 1980s there was growing pressure on lawmakers to rectify what was increasingly seen as a “crisis” in US schools. In response, the Regan administration commissioned a study about the state of public education, and in 1983 published *A Nation At Risk*. The study warned that public education in the US was in a dire situation due to the rising mediocrity of student outcomes that would directly imperil the country’s geopolitical and economic competitiveness (Slater, 2015). Only a top-down reform agenda guided by more rigorous standards – measured by the regular use of standardized testing and enforced through clear incentives and sanctions for schools and teachers – would remediate public education’s precarious state in the US (National Commission on Excellence in Education, 1983).
Uniformly, governors at the time turned to business communities for guidance, who, with an equal uniformity, responded according to the dispositions and knowledge they had available: Schools, it was concluded, should be run more like businesses and hence subjected to hierarchical management, cost-cutting initiatives, standardization, and external control; their success would be best measured through their balance sheets and quantifiable performance (Ravitch, 2002). This paradigm paved the way for the focus of the modern accountability movement on behaviorist logics in which rewards and punishments are seen to motivate changes in performance (Heubert & Hauser, 1999). Under this paradigm, test scores are taken to be directly attributable to educational environments in such a way that poor test performance is seen to be indicative of poor teaching that merits sanctions (Wiliam, 2010), such as the loss of status and possibly the consequent loss of students, staff, and funds, and even the loss of the school altogether if it is converted into a charter (Dworkin, 2005). Although seemingly straightforward, these behaviorist logics rely on the unstated premises that (a) the youth are dangerously ignorant; (b) high-stakes tests are reliable, valid, and appropriate for all students; and (c) the White, middle-class students who consistently serve as the normative reference for these standardized tests are the ideal against which all other students and educational contexts should be measured and toward which they should aspire (Mathison & Ross, 2013). Together, the shifts following A Nation at Risk not only represented a growing national education reform agenda but also a fundamental reconceptualization of the purposes and outcomes of public schooling. Although some educational scholars at the time advocated for alternative conceptions of accountability – such as the proposal by Smith and O’Day (1992-1993) that school reform movements prioritize equality of opportunities to learn challenging content – ultimately the conception of accountability adopted at the national level eschewed equitable inputs in favor of standardized,
quantifiable, and performance-based outputs enforced through punitive consequences (Guiton & Oakes, 1995).

The public appetite for school reform spurred bipartisan cooperation throughout the 1990s leading to the No Child Left Behind (NCLB) act of 2002 (McGuinn, 2006). Under the NCLB, schools were to be primarily evaluated according to performance-based output measures as determined by standardized test results, thus cementing the transition away from input-measures of school quality that had previously defined accountability for almost a century. These output measures were used to evaluate schools and identify low performance, which could then be remedied by additional supports under the School Improvement Grant (SIG) component of the NCLB. Under the Obama Administration, low-performing schools applying for SIG funds were required to implement one of four possible intervention models: transformation (including evaluation, curricular, and structural redesigns), turnaround (including staff layoffs), restart (as a charter or under charter or external management), or closure (Trujillo & Renée, 2015).

Although the NCLB was the reauthorization of the ESEA, unlike the targeted commitments of the ESEA to improving education for students of color and students in poverty specifically, the NCLB sought to reform public education for all students (Cuban, 2004). As such, states and school districts were required to report on the learning of all students as measured by standardized testing, with a special mandate to report on historically marginalized students’ test scores and achievement outcomes as disaggregated subgroups to ensure that these students’ learning received particular attention (Cramer, Little & McHatton, 2018; Fusarelli, 2004). The requirement to report on disaggregated student data was incorporated into the subsequent reauthorization of the ESEA, the Every Student Succeeds Act (ESSA) of 2015. Departing from the NCLB, the ESSA gave states more flexibility in deciding which indicators to
use to measure school success and how much weight to give to each (Callahan & Hopkins, 2017; Darling-Hammond et al, 2016). However, this flexibility has not interrupted the focus on test scores in Colorado, as the Department of Education has opted to continue to primarily rely on outcomes of standardized assessments when evaluating schools (Colorado Department of Education, 2019).

Civil Rights History

Yet the focus on outputs that characterizes contemporary accountability law and policy stands in contradiction to the origins of the accountability movement. Born during the Civil Rights Era, school accountability as established by the ESEA was produced in a social-historical context which included grassroots civil rights organizers who protested racial and linguistic bias in public schools, demanding that school officials be held accountable for the education of students of color as measured both by the output of academic success as well as the input of ending discriminatory practices (Contreras, 2011; Palazzolo, 2013; Roney & Gutierrez, 2019). Despite the history of political and corporate interest in and influence over education reform (Kornhaber, 2004), the need to hold schools accountable for the academic success of historically marginalized populations began during the Civil Rights era of the 1960s and 1970s. However, this movement was not led by politicians but community coalitions which often represented communities of color who were aware that public schools were chronically failing their children (Peck, 2012).

For example, in 1969 students and community members organized in Crystal City, Texas, to fight against a school system which systematically underserved and marginalized Latinx students, ultimately demanding that the school provide bilingual and bicultural education while also improving Latinx representation in the curriculum, teaching staff, administration, and
student activities (Palazzolo, 2013). That same year students and community members organized walkouts in Los Angeles and Denver in response to school systems that both habitually failed to provide Latinx students with equal access to quality education while also overtly discriminating against them. Students in Denver were met by police force when they walked out (Roney & Gutierrez, 2019) and in Los Angeles the community meetings were routinely infiltrated by plainclothes police (Contreras, 2011). Nonetheless these grassroots movements prevailed in not only demanding that their local schools be more responsive to community needs but also achieving actual policy changes which made the schools more accountable for the educational success of historically marginalized students according to community input – not standardized, high-stakes tests. This focus on using accountability as a tool to promote attention to the needs of students of color and bilingual students was again taken up in the 1980s and 1990s when national coalitions such as the National Council of La Raza, the Education Trust, the Citizens’ Commission on Civil Rights, the Center for Law and Education, the Education Equality Project, and the NAACP joined together to organize political and corporate support for accountability, employing conservative ideology and business platforms to successfully argue that improved outcomes for historically marginalized students was both necessary and feasible if federal education reform centered on school accountability and standardization (Rhodes, 2011).

Although in this alignment with conservative and business interests the original grassroots call to focus on inputs was diminished, the disparities were not. Not only were students of color approximately twice as likely to work with ineffective teachers (Darling-Hammond, 1998) but schools that served more low income students and students of color had less access to learning resources like laboratories and computers (Oakes, 1990), a disparity that became exacerbated in high school when such schools chronically lacked advanced placement
courses, tracking students of color and students in poverty into remedial and vocational courses instead (Oakes & Guiton, 1995). These disparities had roots in unequal funding structures across the US which resulted in the wealthiest schools spending up to ten times more than the poorest schools on per pupil student learning, leading to schools that served more students of color lacking textbooks, science labs, licensed teachers, art and music instruction, and functioning bathrooms despite the poorest districts consistently taxing themselves at higher rates than the richest districts (Kozol, 1991). Such disparities of resource investments continue to reverberate today, as high schools with large Black and Latinx student populations less often offer calculus, physics, chemistry, or algebra II as compared to high schools with small Black and Latinx enrollment (Office for Civil Rights, 2016), leaving students in poorer districts with only basic courses structured around rote memorization and vocational tracks while their wealthier peers take classes in foreign language, art, music, technology, and science-based learning (Darling-Hammond, 2013).

Leading up to the passage of the NCLB, educators and researchers increasingly insisted that attention to issues such as access to quality teachers, quality curriculum, and resources be incorporated into any new accountability framework, fearing that failing to do so would only exacerbate the current disparities in outcomes between historically marginalized populations and their dominant-group peers as students and teachers would be expected to perform at higher and higher standards without being given the supports necessary to achieve them (Darling-Hammond, 1998; Guiton & Oakes, 1995). However, the concerns that attention to resources and opportunities were too difficult to measure (McDonnell, 1995), would not necessarily guarantee an increase in achievement (Elmore & Fuhrman, 1995), and were largely irrelevant to learning (Hanushek, 1997) prevailed. In the predecessor to the NCLB, *Goals 2000*, consideration of such
opportunities and resources was optional (Guiton & Oakes, 1995). Under the NCLB, standards for investments in opportunities and resources were totally absent save the requirement that schools employ “qualified teachers,” a term left to be defined by individual states and which, ironically, led to “English Learner” (EL) and immigrant students disproportionately being served by novice teachers in the years after the NCLB was implemented (Dabach, 2015).

As feared, this lack of consideration for how opportunities and resources were invested in schools resulted in schools that serve larger numbers of students of color, students in poverty, and EL students being disproportionately given low accountability ratings which, under the NCLB, were tied to the loss of funds. The loss of students as a direct result of low ratings then further exacerbated the lack of resources that these schools had (Glynn & Waldeck, 2013; Martin, 2012; Martinez-Garcia, LaPrairie & Slate, 2011; McNeil, Coppola, Radigan & Vasquez Heilig, 2008). Worse yet, such low scores prompted many of these schools to narrow the curriculum to only those subjects and skills – including test-taking – that were measured by the standardized tests the NCLB used to evaluate schools. This curriculum narrowing resulted in the loss of challenging curriculum coupled with an incentive for schools to push out low performing students in the hopes of raising test scores (Darling-Hammond, 2007; Vasquez Heilig, Young & Williams, 2012). Together, the loss of funds, loss of students, loss of challenging curriculum, and incentive to push the most vulnerable students out of school resulted in an accountability framework that, in the name of increasing performance, disproportionately punished schools that

---

1 This dissertation will discuss policies and legislation that use deficit language to describe raced, classed, and linguistically marked groups. When referencing those documents, I will endeavor to use the language of the original texts because of the legally-defined nature of the terms. This does not imply that I agree with the deficit language or ideologies behind it. When I describe populations generally and not in relation to specific policies and legislation, I will do so with more inclusive and equity-oriented phrasing. For example, I will use the term “emergent bilingual” except when referencing specific policies and legislation that employ different terminology, such as in this case when “English Learner” references a specific legal designation.
serve historically marginalized students while winnowing the opportunities and resources these students had.

Despite seeking to promote the educational success of historically marginalized students, the accountability movement and the high-stakes, standardized tests which drive it have been broadly criticized for exacerbating rather than remediating the educational inequities that such students face. Emergent bilinguals are disadvantaged by many standardized tests because these tests are normed on monolingual English-speaking populations; since, by definition, emergent bilingual students have not yet mastered English, these high-stakes assessments not only measure content knowledge but also the language skills that such students are necessarily still developing (Abedi, 2004; Menken, 2010; Solórzano, 2008; Tsang, Katz & Stack, 2008). Other historically marginalized students, such as students of color or students in poverty, are likewise disadvantaged by accountability systems which punish schools and teachers for low performance on standardized tests, resulting in de facto policies that encourage low-performing students to drop out or for their teachers to retain them in order to prevent their scores from being recorded (McNeil, Coppola, Radigan & Vasquez Heilig, 2008; Vasquez Heilig & Darling-Hammond, 2008). Such discrepancies make the outcomes of high-stakes standardized tests and consequentially of the accountability frameworks that they inform as much reflections of student demographics as they are of student performance (Glynn & Waldeck, 2013; Martin, 2012; Martinez-Garcia, LaPrairie & Slate, 2011; Strong & Escamilla, 2020), resulting in accountability systems that punish historically marginalized students and their teachers rather than promote learning.
Ideological History

The tension between the grassroots origins of the accountability movement and the contemporary outcomes of those reforms begs the question, why did the latter interpretation of accountability successfully inform national policy reforms while the former did not? In their international comparative study of accountability frameworks Dorn and Ydesen (2014) identified the highly cultural nature of school accountability as it reflects sociopolitical contexts and serves to legitimize certain conceptions of the purpose of education at the expense of others. Although there are hosts of interpretive frameworks available to make meaning of the world (Keane, 2018), the actors and institutions which already possess disproportionate cultural, social, or economic capital are more likely to also have disproportionate access over which interpretive frameworks are used (Bourdieu & Thompson 1991; Fairclough, 1995). This control over ideological resources can be employed in order to promote those worldviews that are most advantageous to the already-powerful by, for example, selectively disseminating the discourses which normalize existing power relations or negative out-group and positive in-group identities (van Dijk, 1993). In this way the NCLB, the ESSA, and any policy text should be seen as both reflecting our social world by employing pre-existing ideas as well as constitutive of it by promoting some ideas at the expense of others (Anderson & Holloway, 2020).

This holds true in the US context as well, which saw the interests of dominant political and economic actors ultimately succeed in defining accountability according to standardized outputs. Seeing an opportunity in the perceived education crisis inspired by the A Nation At Risk report, political and economic elites promoted school reforms that instituted neoliberal logics of individualization by placing the locus of responsibility of institutional success and failure on students and teachers instead of states (Finn, Nybell & Shook, 2010; Wilson, 2018), while
diverting attention away from the local and global contexts in which those students and teachers operate (Burman et al, 2017). This reflected a market rationality that sought to maximize outputs while minimizing investments (Ambrosio, 2013), reconceptualizing the purpose of education away from the production of future citizens and toward the production of future workers (Jenlink, 2016) who, along with teachers, are measured through standardized and thus decontextualized metrics that are presented as stable, unitary, and universally applicable (Gershon, 2016).

Through such quantitatively defined standard metrics like test scores and school ratings, these neoliberal logics were able to then claim that some schools were failing and deserved the punishment of closure (Sunderman, Coghlan & Mintrop, 2017), thereby justifying the privatization of public resources when ‘failing’ schools were consequently converted into private charters (Ambrosio, 2013). Although charters are understandably viewed by some historically marginalized communities as attractive alternatives to public education systems that have failed them, unfortunately they not only do not perform better than public schools (Ravitch, 2010) but they can also further marginalize students, such as in the case of emergent bilingual and special needs students who are enrolled in charters at lower rates due to exclusionary practices and the denial of appropriate services (Shum, 2018). Sadly, it is these very communities that are most negatively impacted, as neoliberal and racial logics converge to justify the transfer of public resources into private hands by framing communities of color as economically irrelevant to economic elites and thus both undeserving of state investments and legitimate targets of disenfranchisement (Lipman, 2013).
Study Site Context

Historical Context

These historical, conceptual, and policy contexts converge at the site of the study, Denver Public Schools (DPS). Understanding the historical context of Denver Public School’s relationship to marginalized student populations demonstrates why it is particularly well-suited for a study regarding the intersection between education policy and provision of educational opportunities and resources with special attention to the needs of emergent bilingual students. In the 1973 Supreme Court case, *Keyes v. School Dist. No. 1*, DPS was ordered to enact racial desegregation in a ruling that was also notable because it confirmed that “hispanics”\(^2\) were an identifiable class for 14th Amendment purposes and thus DPS could no longer argue that a school with a majority African American and Latinx population was desegregated (*Keyes v. School District No. 1*, 1973). In 1980, the Congress of Hispanic Educators (CHE) filed a supplemental complaint based on the Equal Educational Opportunities Act (1974) to argue that “limited-English proficient” students also experienced unequal education. The resulting 1983 District Court case *Keyes v. School Dist. No. 1* found that DPS was obligated to “to take appropriate action to eliminate language barriers which currently prevent a great number of students from participating equally in the educational programs offered by the district” and that “the issues which have been brought before the court by the plaintiff-intervenors [CHE] are part and parcel of the mandate to establish a unitary [desegregated] school system” (*Keyes v. School Dist. No. 1*, 1983).

\(^2\) Although “Latinx” is my preferred term because it is both non-male and non-cis normative, here I use “hispanics” as this was the term used in the Supreme Court ruling.
While not mandating that DPS provide bilingual education, the ruling concluded that providing services to ensure that bilingual students have access to equal education is indistinguishable from desegregation. As a consequence of this ruling, DPS entered into a Consent Decree (CD) in the 1984 *Order Approving Programs for Limited English Proficient Students* which gave the court oversight of DPS’s plans to improve education for emerging bilingual students (*Keyes v. School Dist. No. 1*, 1984). Although DPS was let out of court oversight from the desegregation order in the 1990s, the court has continued oversight of DPS’s provision of services to bilingual students to this day, making this court order the oldest in the country. With CHE as the plaintiff and the Department of Justice as the “plaintiff-intervenor” as of 1999, the most recent iteration of the Consent Decree in 2012 stipulates that DPS engage in systematic tracking of services provided to and outcomes of “English Language Learners” (*Consent Decree of the U.S. District Court, 2012*).

*Accountability Context*

This unique historical context reflects another area in which DPS stands out: It is not only the largest school district in the state, but also the only one that created its own accountability framework rather than use the framework created by the Colorado Department of Education. Since its rollout in 2008, the accountability system designed by the district for its own use, called the School Performance Framework (SPF), had undergone nearly annual revisions due to consistent public backlash over its policies and outcomes (Asmar, 2016b; Asmar, 2017; Asmar, 2019a; Asmar, 2020b) until it was disbanded entirely in 2020 (Asmar, 2020c; Denver Public Schools, n.d. - d). The SPF evaluated schools using different indicators according to school context. Depending on how schools scored across these different indicators, they were given a percentage of points earned out of total points possible, which placed them into one of five color-
coded accountability ratings. Red was the lowest rating possible, followed by Orange, then Yellow, then Green, then Blue, which was the highest (Denver Public Schools (n.d. - d). See Figure 1 for a description of each color-coded SPF rating bracket, the SPF points necessary to achieve each rating bracket, and the district’s description of what each rating bracket indicates regarding school quality.

**Figure 1.**
*Denver Public Schools SPF Color-Coded Rating Brackets Description And Points Cutoffs*

<table>
<thead>
<tr>
<th>Rating (color)</th>
<th>% of Points Earned</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distinguished (blue)</strong></td>
<td>≥79.5%</td>
<td>High-quality school that demonstrates strong results across most areas.</td>
</tr>
<tr>
<td><strong>Meets Expectations (green)</strong></td>
<td>≥50.5% and &lt;79.5%</td>
<td>Quality school that demonstrates good results in many areas, with a few areas in need of some improvement.</td>
</tr>
<tr>
<td><strong>Accredited on Watch (yellow)</strong></td>
<td>≥39.5% and &lt;50.5%</td>
<td>School that demonstrates results in some areas and/or has several areas in need of improvement.</td>
</tr>
<tr>
<td><strong>Accredited on Priority Watch (orange)</strong></td>
<td>≥33.5% and &lt;39.5%</td>
<td>School that demonstrates low results overall and/or has multiple areas in need of significant improvement.</td>
</tr>
<tr>
<td><strong>Accredited on Probation (red)</strong></td>
<td>&lt;33.5%</td>
<td>School that demonstrates minimal results overall and has many areas in need of significant improvement.</td>
</tr>
</tbody>
</table>

The bulk of SPF scores were determined by students’ performance on annual state-administered standardized assessments (Denver Public Schools, n.d. - e). Consistent with federal requirements, these test scores were used to calculate both single-year snapshots of student performance, called Status, as well as year-to-year measures of advances in learning reported as student Growth (Asmar, 2019a), with Growth being weighted more heavily than Status (Asmar, 2017). In addition, in the 2016-2017 academic year the district implemented an Equity Indicator, which described the degree of performance differentials between dominant and marginalized students’ standardized test outcomes wherein schools with large “academic gaps” were prohibited from receiving the highest SPF rating despite their scores on all other indicators (Asmar, 2016b). A small percentage of schools’ SPF scores were also derived from the results of the Parent and Student Satisfaction surveys and, for high schools, graduation rates and post-secondary readiness (Denver Public Schools, n.d. - e).

The SPF operated under a behaviorist perspective in which rewards and punishments are seen to motivate schools and teachers to perform differently (Dworkin, 2005). In DPS, the SPF was intended to reward high performing schools with publicly available desirable SPF ratings while identifying low-performing schools for extra supports and, if improvements were not achieved, negative reinforcements (Asmar, 2017). Negative reinforcements included reduced teacher pay, mandated improvement plans, possible restart or closure, and publicly available low ratings which reduced enrollment and funding. These negative consequences mirrored those mandated by the federal government under the Obama Administration in which low performing schools seeking federal grants were required to implement interventions of either transformation, turnaround, restart, or closure (Trujillo & Renée, 2015).
The publicly available SPF ratings translated into more families choosing Blue or Green rated schools than Red schools under the district’s universal school choice model, thereby reducing low-performing schools’ enrollment and the funding attached to it (Asmar, 2019a) and jeopardizing these schools’ ability to fund teachers and programming (Asmar, 2019b). SPF ratings also impacted teacher pay, as the district used an incentives-based system (Asmar, 2016c). Although the district shifted its policy regarding school restarts and closures several times in the last decade, Red and Orange status triggered review and intervention in both of the most recent policies. Beginning in 2015, the district adopted the School Performance Compact (Denver Public Schools, n.d. - a), which initially used SPF scores to identify the lowest performing schools for review and potential closure or restart (Asmar, 2016a), with schools scoring Red two years in a row or a Red or Orange rating in the two years preceding a Red rating automatically being slated for restart or closure if student progress was insufficient (Asmar, 2018; Denver Public Schools, 2018). In 2018, that policy was revised to mandate that all schools with two years of Red SPF ratings or an Orange SPF rating followed by a Red rating needing to submit improvement plans, which would then be reviewed by a committee before an intervention – ranging from one or two years of monitoring to restart or closure – was suggested to and voted upon by the school board (Asmar, 2018), although this policy was also suspended due to community concerns in the 2018-2019 academic year (Denver Public Schools, n.d. - a).

However, it must be noted that district-provided information about the district’s policies regarding interventions following low SPF ratings was unclear, as the district describes its policy in vague terms such as, “When schools do not meet expectations for academic growth and achievement on the School Performance Framework, DPS provides intensive support to help them get back on track…. Although a restart, turnaround or closure is never an ideal outcome, it
is sometimes necessary” (Denver Public Schools, n.d. – a), and, “If schools are not able to show improvement after significant support efforts over time, DPS believes that the students served by these schools deserve a major change in their learning environment” (Denver Public Schools, 2018). Neither of these statements nor any publicly available information from the district specifies exactly what kinds of supports are provided, what amount of time schools have to show improvement, what constitutes significant improvement, what “major change in learning environments” entails, what “restart” entails, or how and under what circumstances schools are closed. Upon request for clarification, the district representative for accountability was unresponsive.

Research Context

Taken together, the federal mandate for disaggregated reporting found in the ESSA and the local mandate for disaggregated reporting found in the Consent Decree result in a unique accountability context in which DPS is required to provide exceptional amounts of accountability data due to its historical struggle to provide equitable education to historically marginalized populations and English Learners in particular. Currently, DPS serves over 90,000 students, most of whom are Latinx (approximately 64%) and approximately a third of whom (37%) are labeled as “English Learners.” The student population of DPS is also 13% African American and 23% White, with 67% of all students qualifying for Free and Reduced Lunch services, a proxy for socioeconomic status, and 11% of students classified as receiving Special Education services. While these characteristics are not uncommon in the US, they do result in the district serving primarily students of color and students in poverty, with over one in three students requiring special language interventions, tracking, and reporting per the CD.
Despite the flexibility of the ESSA and the commitment to serving bilingual students and students of color expressed by the current superintendent, Alex Marrero, DPS has struggled to implement school accountability that equitably measures student learning. A recent study by Strong and Escamilla (2020) found that SPF school ratings had statistically significant relationships to English Learners’ levels of English proficiency as measured by student WIDA ACCESS scores, making the SPF not only a reflection of student learning but also of student demographics. Since neither schools nor teachers can control what kinds of students they serve, an accountability framework that penalizes them for serving historically marginalized populations threatens to stigmatize these students as educators may falsely conclude that the negative outcomes of low SPF ratings – including district intervention, loss of students and funds, and school closure – are attributable to the students themselves rather than a faulty accountability system (Strong & Escamilla, 2020).

**Theoretical Framework**

*Critical Race Theory*

The ways that accountability policies both purport to serve historically marginalized students while also resulting in outcomes in which the schools that serve these students are disadvantaged and further marginalized is predictable if one employs a Critical Race Theory (CRT) lens, which understands schools to be sites of social reproduction that are socially- and historically-situated. This perspective allows us to connect current practices in education to broader patterns of institutional power disparities across society. Critical Race Theory is both a theoretical and methodological framework with origins in critical legal studies (Bell, 1980; Matsuda, 1993). CRT posits that, rather than a temporary or isolated aberration, racism is endemic throughout US society and institutions (Russell, 1992) and thus must serve as a focal
point into any social science research (Howard & Navarro, 2016). By foregrounding race, research is thus able to highlight the processes by which racialization reproduces inequality such as through public education policy (Gillborn, Warmington & Demack, 2018). Racialization does not reflect inherent, essential differences between individuals (Bonilla-Silva & Zuberi, 2008) but rather is a social construct whose fluid boundaries shift in the service of maintaining a white supremacist social order (Roediger, 2005). Race as a social construct nonetheless has real, material consequences, such as regarding the allocations of school resources and opportunities (Ladson-Billings & Tate, 1995), the right to property (Harris, 1993), and even language status designations (Rosa, 2016). Furthermore, a tenet of CRT is that, although racialization is endemic throughout US society, it interacts with other forms of oppression and power to create unique intersectional identity categories that must be understood holistically rather than reduced to their constituent parts (Crenshaw, 1991). In this way, students’ race interacts with students’ economic status, gender, language practice, and cultures to create specific contexts in which students’ opportunities are afforded or denied (Howard & Navarro, 2016).

CRT explicitly calls attention to the fact that students of color, students in poverty, and emergent bilingual students are poorly served by public schools, often as the result of policies, legislation, and practices that systematically disenfranchise such students and the communities from which they come (Baker & Wright, 2017; Donato, 1997; Donato & Hanson, 2012; Flores, 2005; Kozol, 1991; Leonardo, 2015; Menchaca, 1993; San Miguel & Donato, 2010; Santa Ana, 2004). This patterned disenfranchisement can be seen in part as stemming from disparities in school contexts and investments, making the “achievement gap” more accurately understood as an “opportunity gap” (Ladson-Billings, 2013b). Although such systemic educational disenfranchisement has contributed to the inequalities which accountability reforms purport to
address, CRT scholarship clarifies this seemingly contradictory dynamic through the concept of interest convergence, or the way that institutional policies like accountability are justified through liberal, race-neutral, or even social justice frameworks but in reality serve to further the interests of dominant-group members while obscuring how such policies are actually self-serving (Bell, 1980), although some scholars contend that the interest convergence concept centers whiteness and dominant group members at the expense of explicit attention to the needs and interests of communities of color (Garces, Ishimaru & Takahashi, 2017).

**QuantCrit**

Building off of Critical Race Theory, QuantCrit posits that, due to the problematic history and uses of demographic statistics, researchers employing these methods must do so carefully and only for social justice purposes lest the research inadvertently perpetuates the white supremacist status quo (Gillborn, Warmington & Demack, 2018). The tenets of QuantCrit, discussed below, can be summarized as: (a) racism is central to US society; (b) quantitative data are neither objective nor politically neutral but imbued with social and research bias, like any kind of data; (c) likewise, racial categories are not objective, stable, or natural but social constructions; (d) quantitative data, like all data, require interpretation, the act of which is imbued with researcher bias, and thus numbers should not be taken to ‘speak for themselves’; and (e) because of the ideological and political nature of quantitative research, like all research, in addition to its problematic history of being used as justification for a white supremacists and oppressive social order, researchers using quantitative data must do so with explicitly antiracist designs and purposes (Gillborn, Warmington & Demack, 2018).

It must be noted that using statistical methods to study the relationship between student demographics and learning outcomes is highly problematic, especially considering one of the
demographic variables of interest in this study – student race. Born during a time of overt racism, statistical analysis of racial demographics was developed as a part of the eugenics movement, which sought scientific justification for white supremacy through the “objective” measurement of the “inferiority” of people of color that “necessitated” a racist social order (Zuberi, 2001). In addition, racial categories themselves are highly arbitrary and unstable due to the socially constructed nature of race, making racial categorizations fluid and ideational rather than a discrete, consistent marker of biological difference (Bonilla-Silva & Zuberi, 2008). The use of racial categories to define populations is thus both error prone and easily co-opted by white supremacist interpretations of racial groups’ different outcomes as implications of inherent racial difference and racial superiority/inferiority (Gillborn, Warmington & Demack, 2018).

These concerns beg the question, why measure racial groups at all? Although race is a social construct, the ways individuals are racialized in society has very real material consequences. Because public schools, like many institutions in the US, operate as systems of social reproduction (Bourdieu & Thompson, 1991; Covarrubias, 2011), and the historical foundation in this country is built on white supremacy (Russell, 1992), public education in this country often reflects and reproduces white supremacist social orders (Leonardo, 2015; Yosso, 2002). For example, in a review of all 50 states’ elementary and secondary standards, Sabzalian, Shear and Snyder (2021) used descriptive statistics guided by QuantCrit and TribCrit to reveal that over half of states have little or no discussion of tribal sovereignty anywhere in their K-12 standards. For reasons like these, it is imperative that educational researchers are attentive to issues of race in public schools without reifying racist narratives of inherent racial difference (Crawford, Demack, Gillborn & Warmington, 2019).
In order to do this in quantitative research, scholars can adopt a critical stance toward statistical methods and data (Stage, 2007). QuantCrit scholars note that although quantitative data are often taken by academics, policymakers, and the public as somehow more objective than qualitative data, they are nonetheless still subject to a host of researcher biases at every step of the research design, implementation, and interpretation: Which questions are asked, how populations are defined and measured, and the ways that relationships between population groups and social outcomes are interpreted are all reflections of researcher choice and hence potentially researcher bias (Crawford, Demack, Gillborn & Warmington, 2019; Stage, 2007; Suzuki, Morris & Johnson, 2021). Thus, quantitative data is no more apolitical, objective, or value-free than qualitative data and need to be understood for their subjective, political nature (Bonilla-Silva & Zuberi, 2008).

Although a disperse framework, QuantCrit is guided by several positions derived from Critical Race Theory such as the understanding that white supremacy is endemic in US society (Bell, 1992) and thus necessarily constitutes the context of all our social practices and hence research foci, making all research essentially political (DeCuir-Gunby & Thandeka, 2019). Critical scholars, then, must be explicitly dedicated to antiracist research approaches lest our findings by default perpetuate the white supremacist status quo (Garcia, López & Vélez, 2018).

**Conceptual Framework**

Taken together, Critical Race Theory and QuantCrit frameworks are useful tools to highlight the mechanisms by which school accountability functions to both further disenfranchise historically marginalized communities as well as legitimate that marginalization. These frameworks call on scholars to be attentive to the processes of marginalization, such as through the role of the privatization of public resources as in the example of charters. They also
call on scholars to interrogate the relationship between racialized populations and disparate investments and outcomes that purportedly justify such privatization. This dissertation drew on the explicitly antiracist foundations of both of these frameworks in its intention and design, as it sought to challenge deficit views of historically marginalized communities by highlighting patterns of systemic oppression (Yosso, 2002), such as how institutional processes like accountability reproduce systemic inequalities through the disparate school contexts in which racialized, classed, and linguistically marked students find themselves and that result in disparities in these groups’ academic outcomes. In addition, these two critical theoretical frameworks were employed to conceptualize how variables extrinsic to the SPF accountability framework that reflect institutional patterns of marginalization might impact accountability outcomes, thereby guiding the selection of quantitative variables for analysis. Specifically, this study explored the intersections between SPF accountability outcomes and (a) student demographics, (b) school contexts identified in previous research that reflect student demographics (teacher quality, discipline, charter status, and enrollment), and (c) characteristics and services relevant to English Learners specifically. The last set of variables reflects both the CRT concept of intersectional identities of racialized students in which their language and class status contribute to unique social locations and thus avenues of marginalization as well as the specific context of the study site, Denver Public Schools, whose struggles to adequately serve English Learners resulted in the intervention of the Department of Justice as discussed in the above section. The rationale for selecting these variables for analysis is presented below.

*Student Demographics*

This study included variables describing student populations of students of color, students receiving Free and Reduced Lunch services (a proxy for classed status via income), and students
with the English Learner label due to research which shows that racialized, classed, and linguistically marked populations are often denied the academic services and resources necessary for school success (Darling-Hammond, 2004; Martin, 2012; Wu, 2013). Critical Race Theory and QuantCrit describe this maldistribution of resources and opportunities as reflections of educational failures rather than students and families (Morris & Parker, 2013; Ramlackhan & Wang, 2021). Together, this study included student demographic variables to investigate whether accountability outcomes mirrored student demographics while framing any disproportionality under a CRT and QuantCrit lens as reflections of systems rather than communities.

An additional area of systemic disproportionality that has been shown to negatively impact historically marginalized students deals with placement and referrals to either Gifted and Talented (GT) programs and Special Education (SPED) services. Black and Latinx students are more likely to be referred for Special Education services than their White peers (Tenenbaum & Ruck, 2007) even within income categories, indicating that race rather than socioeconomic status alone is a stronger predictor of disproportionate SPED placements (Grindal, Schifter, Schwartz & Hehir, 2019). Similarly, English Learner students, taken to be students of color, are more likely to not only be labeled as having learning disabilities but also mental retardation relative to their White peers (Sullivan, 2011).

The potential institutional biases against students of color that result in these disproportionalities are also evident in the underrepresentation of the same students in GT programs (Grissom & Redding, 2015). The propensity to under-identify the talents of historically marginalized students was exemplified when a large school district in Florida transitioned from a system for identifying students for GT based on teacher referrals to one based on universal screening where, without changing standards for entry into GT programs, the
district saw enormous increases in girls, students of color, students in poverty, and English Learner students all qualifying for placement (Card & Giuliano, 2016). The role of teacher bias in failing to identify the talents and strengths of historically marginalized students as evidenced in these studies might explain the national patterns of disproportionality in GT programs wherein Black and Latinx students, who represent 42% of enrollment in schools that have GT programs, only represent 28% of GT participants, a similar dynamic to that of English Learners who likewise make up 11% of students by only 3% of GT participants (US Commission on Civil Rights, 2018).

The overrepresentation of historically marginalized students in SPED programs and the underrepresentation of these same students in GT programs represent a failure to provide these students with appropriate educational services as placement decisions are based on students’ racial, class, and language statuses rather than learning needs. As such, this study operationalized the Critical Race and QuantCrit frameworks which highlight these patterned disparities through the inclusion of variables describing the rates of GT and SPED participation in each school in addition to variables describing rates of students classified as students of color, students receiving Free and Reduced Lunch services, and English Learner students.

School Contexts

Another set of study variables focused on school contexts, including enrollment, student-teacher ratios, charter status, discipline rates, and teacher quality. Students in poverty, students of color, and emergent bilingual students are less likely to work with highly qualified teachers (Darling-Hammond, 2004; Goldhaber, Lavery, & Theobald, 2015; Lankford, Loeb & Wyckoff, 2002), whether that be measured through years of experience (Clotfelter, Ladd & Vigdor, 2005; Dabach, 2015), teacher effectiveness ratings (Borman & Kimball, 2005), or teachers having
majored or minored in the content area if teaching middle or high school (Jerald & Ingersoll, 2002; Peske & Haycock, 2006). In fact, an international comparative study found that the US ranked fourth of 46 countries in disparate access to “quality” math teachers between high- and low-SES students (Akiba, 2007). Teacher qualifications – especially certification and education in the content area – have been found to be more impactful than student demographics, class size, teacher salaries, and general school funding in affecting student achievement (Darling-Hammond, 2000), making the disparate access to quality teachers across raced, classed, and linguistically-marked students especially troubling. Strong and Escamilla (2020) found that the schools serving larger proportions of students designated as English Learners had smaller average proportions of their teaching body that were designated as “Fully Qualified” to teach bilingual students relative to schools that served student populations that were Whiter, wealthier, and more composed of English-monolingual and English-proficient students. Applying a CRT to this issue, disparate access to “Fully Qualified” teachers is a metric that likely represents mechanisms by which the district denies full investments of resources and opportunities to raced, classed, and linguistically marked populations, resulting in disparate accountability outcomes which in turn exasperate the lack of equitable access. For this reason, in this study CRT and QuantCrit frameworks are operationalized through the inclusion of a variable describing the percentage of “Fully Qualified” teachers at each school.

Variables describing disciplinary environments in schools were also included in the study, as the race and socioeconomic status of students has been found to predict rates of disciplinary referrals (Bryan, Day-Vines, Griffin & Moore-Thomas, 2012; Skiba, Chung, Trachok, Baker, Sheya & Hughes, 2014), with Black students overrepresented in both K-12 and also tragically preschool suspensions and expulsions (US Commission on Civil Rights, 2018).
Teachers have been found to not only direct more positive speech at White students, but also recommend them less often for behavior referrals when compared to Black and Latinx students (Tenenbaum & Ruck, 2007) with White teachers more likely to interpret their Black students’ behaviors as disruptive (Bates & Glick, 2013; Wright, 2015). Such dynamics result in students of color being overrepresented in disciplinary actions that remove them from school all together, such as out of school suspensions (Anyon, Wiley, Samimi & Trujillo, 2021). In the study site, Denver Public Schools, these trends also hold, with Black students being overrepresented in law enforcement referrals, tickets, and arrests (Asmar, 2020a). This study posits that disproportionality in discipline is an additional quantitative measure of how disparate school contexts impact historically marginalized students’ educational experiences and consequentially accountability. In this way, this study operationalized CRT and QuantCrit frameworks through the inclusion of variables describing the rate of disciplinary (a) incidents, (b) actions, and (c) actions that resulted in loss of instructional time.

Another school context variable included in the study described whether a school was a charter or district-run. Charters have grown in popularity due to the perception that they offer market-based alternatives to the inefficient bureaucracy of public schools, as their focus on pleasing their clients, who are likewise free to choose the schools with the best results, will spur innovative improvements in organization and learning (Chubb & Moe, 2011). By being driven by competition, the theory holds that charter schools will achieve superior outcomes and be more responsive to community needs, as those that fail to do so will also fail to attract the requisite families needed to operate and will be forced to close (Howe, Eisenhart & Betebenner, 2002). Because of this perception, charter schools are seen as attractive alternatives to district-run schools that suffer from low accountability ratings, and through the “Call for New Quality
Schools” process Denver Public Schools allows for new schools to be created to replace those that are low-performing (Denver Public Schools, 2018). However, whether charters actually result in greater student achievement is unclear, as the corporate and market logics which undergird charters substitute attention to the impact of race- and class-based inequities in producing disparate learning outcomes with solutions grounded in competition, standardization, and resource management, which potentially exacerbate these inequities as schools are free to reject the highest-needs students (Kantor & Lowe, 2016). This dynamic reinforces the disparate opportunities and resources afforded raced and classed families, who can be excluded from high performing charters through exclusionary enrollment practices and then blamed under the market logics of individual responsibility for participating in low-performing schools (Howe, Eisenhart & Betebenner, 2002; Lipman, 2013). For these reasons, CRT and QuantCrit were operationalized in this study through the inclusion of data describing the charter status of schools, as this status has special implications both for accountability outcomes as well as how well historically marginalized populations are served in the district.

Finally, this study also examined school context variables describing student population sizes as an absolute value of total enrollment and a relative value as the ratio of students to teachers. Data regarding school enrollment size was included because research has found that it can impact relative disadvantages of students of color in public schools, with larger schools having greater structural disadvantages for students of color compared to White students (Fitzgerald, Gordon, Canty, Stitt, Onwuegbuzie & Frels, 2013). In addition, enrollment size is currently being considered as a metric to use in decisions regarding whether to close or consolidate schools at the study site, Denver Public Schools (Asmar, 2021), making it not only a
timely variable for analysis but potentially an avenue that can lead to the same end of school closure that parallels the outcomes of accountability ratings.

Similarly, student-teacher ratios have been used in previous research as a factor that can impact student learning independent of the other variables used in this study (Driscoll, Halcoussis & Svorny, 2003; Powers, 2003; Wu, 2013), making it a necessary control and example of school context factors external to metrics measured by the accountability framework used in Denver Public Schools. Because CRT and QuantCrit prioritize examinations of systemic mechanisms by which institutional oppression is reproduced and legitimized, this study operationalized these theoretical frameworks through the inclusion of these student population size school context variables in addition to those of charter status, disciplinary environments, and teacher quality, as all have been found to relate to historically marginalized student populations’ disparate access to high quality education.

**English Learner Characteristics and Services**

Emergent bilingual students have a unique historical context of being marginalized and racialized. Thus it is critical to center their skills and learning needs rather than applying a color-blind lens (Bonilla-Silva, 2006) that would treat these students and their history and needs as indistinguishable from dominant group students. As such, this study included metrics to evaluate the characteristics and educational services unique to emergent bilingual students through variables describing (a) placement of ELs in programs to develop English, (b) the parent preferences for such program placement, (c) the rates at which English Learner students are redesignated, exited, and re-entered into English Learner status, (d) the level of English proficiency of students, (e) the language status of English Learner students, and (f) the representation of English Learners in Special Education (SPED) and Gifted and Talented (GT)
programs. This is because if these students learn in environments in which their bilingualism is positioned as a deficiency to be overcome rather than assets to be embraced or is otherwise subtractive through assimilationist ideologies and monolingual normativity (Valenzuela, 1999), then these students will experience diminished school quality despite their school contexts and school demographics.

Research has shown that when emergent bilingual students have the opportunity to develop their bilingualism and biliteracy, they do better in math, reading, and even English (Ramírez, 1992; Thomas & Collier, 1997), making English-only or transitional bilingual education ineffective if educators are seriously committed to promote their academic achievement (Rolstad, Mahoney & Glass, 2005). However, a 2010-2011 representative sample of kindergarten students participating in English Learner programs found that only 8% of such students participated in programs designed to develop bilingualism (Redford, 2018). Instead, most emergent bilingual students are denied these opportunities and advantages. Making matters worse, when schools receive low accountability ratings the curriculum is often further narrowed to only teaching those subjects and skills that will be reflected in accountability measures (Diamond & Spillane, 2004), which can even lead to the loss of otherwise successful bilingual education programs (Menken & Solorza, 2014).

The loss of bilingual programming due to accountability frameworks that do not value and hence do not measure bilingualism is an example of how language policies historically and currently act as educational gatekeepers in the service reproducing power differentials in society (Tollefson & Tsui, 2014). In the nineteenth century, Native American children were forcibly interned in boarding schools and prohibited from speaking in their home languages through a violent policy of linguistic and cultural assimilation (Wiese & Garcia, 1998). Today, anti-
immigrant and assimilationist ideologies are still evident in English-only programs and accountability paradigms that implicitly position students’ home languages and cultures as obstacles to overcome rather than assets (Baker & Wright, 2017; Black, 2006; Wiley & Wright, 2004), despite research showing the academic, interpersonal, and cognitive benefits of bilingualism (Bialystok, Craik, Green, & Gollan, 2009; Dorner, Orellana & Li-Grining, 2007; Martínez, 2010). Such ideologies are also evident in the invalid yet ubiquitous practice of using monolingual-normed tests to assess bilingual students, which have led to the misidentification of students as in need of remediation or even as lacking language (Hopewell & Escamilla, 2014; MacSwan, 2005).

Because of this unique historical and contemporary context, the CRT attention to intersectional mechanisms of oppression was operationalized in this study through the inclusion of variables that described both the variation of English Learners’ language needs as well as the ways that different school contexts supported those needs or failed to do so. These variables described both what kinds of language support programs English Learners were placed in as well as their families’ preferences for language support programs. These data were complemented by variables describing the rates at which English Learners were redesignated from, exited from, and re-entered into English Learner status, as holding this label has been shown to reduce students’ access to challenging curriculum (Brooks, 2020) and negatively impact students’ learning outcomes (Kim, 2017). In addition, this study examined variables describing English Learners’ level of English proficiency according to WIDA ACCESS scores, with different levels of English proficiency indicating distinct student needs in addition to being statistically significant predictors of accountability outcomes (Strong & Escamilla, 2020).
Variables describing English Learner characteristics included the rates at which English Learners were placed in SPED and GT programs, as English Learners are often overrepresented in SPED (Sullivan, 2011) and underrepresented in GT programming (US Commission on Civil Rights, 2018). Additionally, this study included a variable to describe if English Learners were Spanish-speakers. This is due to evidence that raciolinguistic ideologies that position racialized Spanish-speaking bilinguals as particularly linguistically deficient as a means of legitimizing white supremacist racial hierarchies (Hill, 2009; Rosa & Flores, 2017), which potentially parallels the role of accountability outcomes in legitimizing the maldistribution of opportunities and resources across schools (Lipman, 2013).

In order to operationalize the CRT focus on how institutional and ideological mechanisms disenfranchise students with complex intersectional identities, this study also included analyses of disproportional rates of English Learner representation in SPED and GT programs and the disparate outcomes and school contexts of Spanish-speaking English Learners. Together with quantitative data describing language service program placement, family desires for language services, rates of redesignations, exits, and re-entry into the English Learner label, and level of English proficiency, this dissertation explored variables specific to the characteristics and services unique to English Learners.

Summary

By operationalizing CRT and QuantCrit to include the learning contexts and needs specific to historically marginalized students, this research project sought to decenter normative whiteness, monolingualism, and middle-classness while drawing attention to potential areas for accountability focus that might be more effective in explaining and overcoming disparate achievement outcomes.
Purpose of Study

This research project applied a Critical Race and QuantCrit lens to explore this unique historical and accountability context of Denver Public Schools. This study aims to explore the relationship of non-achievement metrics like student demographics, school contexts, and English Learner characteristics and services to other metrics used in the SPF. Research has shown that these broader factors impact learning outcomes (Guiton & Oakes, 1995; Teddlie, Stringfield & Reynolds, 2002; Wang, 1998; Wu, 2013), yet by relying just on the SPF, DPS leaders and educators were not able to consider them. Specifically, this project seeks to expand our understanding of how and why accountability is defined by school districts. This study is particularly relevant as school districts around the nation are rethinking their accountability systems. Capitalizing on the expansive accountability reporting available in DPS and including metrics describing student demographic, school contexts, and English Learner characteristics and services that research has shown impact historically marginalized populations’ educational experiences can hopefully broaden policymakers' understanding of schools' unique contexts and needs.

Critical Race and QuantCrit frameworks alert us to the need to examine the ways that institutional policies like accountability result in the disenfranchisement of historically marginalized communities, especially through ostensibly race-neutral policies like accountability (Gillborn, 2005). Taken together, this study sought to identify the relationships between school SPF accountability outcomes and the student demographics, school contexts, and English Learner characteristics and services that are not measured by the SPF. This approach offers to potentially highlight the ways in which the SPF erroneously measures – and holds teachers and
schools accountable for – non-academic, contextual, and demographic variables of schools instead of only student learning and school quality.

This investigation also offers to highlight the ability of the SPF to achieve its goal of promoting improved student learning and school quality. If the SPF accountability framework is effective in promoting school success as revealed through SPF ratings, then trends should indicate that schools gain or stay in the lowest categories briefly as the accountability consequences of low status result in improved performance. Conversely, as a result of accountability consequences that encourage high performance, over time schools should increasingly gain and maintain designations in the high performing rating designations. Finally, not only should an effective accountability framework result in brief low ratings designations and greater rates of entering into and remaining in high ratings designations, but it should also only reflect school success rather than student demographics. For this reason, if the SPF accountability framework only reflects school success, then the characteristics of the schools in the extremes of the highest and lowest SPF ratings should be approximately similar. As such, this study centers the student demographic, school context, and EL characteristic and services variables extrinsic to the SPF, as an unbiased accountability framework should not reflect any of these non-achievement metrics, and these non-achievement metrics should likewise have no impact on schools’ SPF ratings.

Research Questions

In order to investigate the relationship between accountability outcomes and student demographics, EL characteristics and services, school contexts, this study addressed the following research questions:
1. What are the student demographics, EL characteristics and services, and school contexts per SPF rating bracket?

2. At what rate do schools remain in, enter into, or exit the most extreme SPF ratings statuses of Intervention vs. Blue, and what are the student demographics, EL characteristics and services, and school contexts in them?

3. What are the student demographics, EL characteristics and services, and school contexts per charters and district-run schools?

4. Do student demographics predict SPF scores?

5. Do student demographics predict SPF outcomes?

**Significance of the Research**

I posit that it is only through understanding a problem that one can work toward a solution. In the same way, it is only through better understanding the institutional mechanisms through which disenfranchisement occurs that targeted policy solutions can be crafted and implemented. Better understanding of how student demographics, school contexts, and accountability outcomes interact with attention to student-specific needs and outcomes can potentially help policymakers reframe how accountability is conceptualized, institutionalized, and enacted in order to confront systemic disparities in accountability outcomes that punish historically marginalized students and the schools that serve them. Empirical data regarding the ways that the most recent accountability framework used by Denver Public Schools impacted historically marginalized populations can help district leaders design better, more equitable accountability policies and practices that measure and provide the students-specific resources that different historically marginalized demographics such as emergent bilinguales deserve. Given the country’s long history of educational disenfranchisement of raced, classed, and linguistically-
marked students, research like this study that seeks to identify policy mechanisms of marginalization in order to craft improved services and outcomes for such students is timely as ever.
**Literature Review**

The literature review follows three aspects of extant research to explore studies regarding: (a) the efficacy, validity, and utility of accountability policies, (b) the efficacy and outcomes of accountability policies for historically marginalized students in particular, with a special focus on emergent bilingual students, and (c) QuantCrit approaches to understanding school environments and outcomes for historically marginalized populations. To conclude, the connection between previous research explored here and the current dissertation is discussed.

The search for literature was conducted using ERIC (Proquest) and Education Full Text (EBSCO) in addition to bibliographic chaining. Search terms for the first two sections (regarding the efficacy, validity, and utility of accountability policies generally and for historically marginalized students specifically) were: “school accountability” in combination with “validity” or “efficacy” or “outcomes.” These results were then coded to indicate when the studies described general accountability research and when they described issues pertinent to historically marginalized students, with those about emergent bilingual students being sub-coded as a distinct category. Search terms for the third section (regarding QuantCrit studies) were: “school accountability” in combination with (a) “historically marginalized students” or “student demographics,” and (b) “QuantCrit” or “critical race theory.”

Results were refined in three ways: (a) to only books and articles (excluding dissertations, opinion pieces, and other types of media); (b) to only publications after 2000 to represent the current accountability era characterized by the passage of the No Child Left Behind act; and (c) to only those pertinent to the US context, although if studies also discussed international contexts in addition to the US, as in the case of the Williams (2010) piece, they were also included.
Research Regarding the Validity of Accountability Policies

The most recent federal accountability law, the Every Child Succeeds Act (ESEA) of 2015, allows individual states leeway to implement accountability policies, resulting in a patchwork of systems in which states use different indicators and different weights (or some not at all) to measure different constructs (Darling-Hammond et al, 2016). This flexibility came on the heels of an already lenient system which was prone to inconsistency such as the utilization of divergent cutoff points, tests used, growth scales, and the incorporation of non-academic factors.

A comparative cross-state analysis (Martin, Sargrad, Batel & Center for American Progress, 2016) concluded that cross-state variation was so diverse that a child could very easily be considered proficient in a subject area in one state only to find that in the next she is below average. Similarly, using a path analysis of relationships between policy inputs, outcomes, and contexts of all 50 US states Lee (2010) found that, because states had wide latitude in implementing federal standards, some opted to manipulate their own standards frameworks in order to artificially inflate their scores and delay implementation of federal objectives entirely so as to avoid the negative repercussions related to under-performance (a rational choice when accountability standards demanded increased performance without providing increased supports). Such “gaming” of the accountability system was also found by Vasquez Heilig (2011), whose longitudinal descriptive cohort progression analysis of 45,000 students investigated accountability reporting outcomes, finding that the publicly reported graduation rates were mathematically impossible.

Beyond the manipulation of accountability frameworks by districts and states, research has questioned whether even faithfully implemented accountability systems actually reflect student learning or school success. In a qualitative comparison of accountability ratings systems,
Murray and Howe (2017) concluded that systems which report single metrics of school success like a letter grade are unlikely to accurately describe school quality or motivate the very improvements upon which the accountability system is premised. This is because such oversimplified ratings don’t actually reflect differences in performance. A quantitative study using multilevel modeling and ANOVAs conducted by Adams, Forysth, Ware, Mwavita, Barnes, and Khojasteh (2016) revealed that students from the schools with the highest ratings did not have statistically significant higher reading or math outcomes than students in schools with lower ratings. More striking, the highest-rated schools were also home to the greatest achievement gaps between students receiving Free and Reduced Lunch services (FRL) and students of color as compared to their non-FRL and White counterparts. This is due in part to the fact that FRL students and students of color in schools with the lowest ratings actually had higher average reading and math performance than FRL students and students of color in the highest-rated schools (Adams, Forysth, Ware, Mwavita, Barnes & Khojasteh, 2016).

Disaggregating outcomes by student demographics could help address the misleading nature of single-metric accountability ratings. For example, Glynn and Waldeck (2013) found in their comparative analysis of SchoolDigger school ratings in four states that not only did single-metrics ratings often obscure such achievement gaps but they also represented variation in student outcomes that were not statistically significant between one rating category and the next. However, disaggregation is not a panacea to the unreliability of accountability scores, as highlighting performance by student subgroups introduces new problems. Vasquez Heilig, Young, and Williams (2012) found in their qualitative study using focus groups and interviews that teachers and administrators saw “at-risk” students’ disaggregated lower test scores and consequently interpreted these students as threats to the school’s accountability rating. Similarly,
in a mixed methods case study of a Latino-majority high school over seven years, McNeil, Coppola, Radigan, and Vasquez Heilig (2008) found that disaggregating outcomes by student demographic led to the view that historically marginalized students were not pupils to be taught but liabilities who imperiled accountability ratings.

The accuracy of the tests used to determine accountability ratings has also been questioned as many standardized assessments reflect constructs beyond student content learning. For example, Spees, Potochnick, and Perreira (2016) used regressions to evaluate the relationship between individual-level eighth grade scores on the National Association of Educational Progress, student demographics, and contextual factors including type of city. They found that, in the case of emergent bilingual students, students’ scores reflected whether they lived in new or established immigrant communities, demonstrating how variables far removed from the quality of instruction can impact test outcomes and consequently accountability ratings.

In a review of literature regarding the validity of teacher evaluation instruments used for teachers of emergent bilingual students, Turkan and Buzick (2016) found that because "there is no uniform definition of necessary teaching knowledge and skills to be effective teachers of ELLs [or English Language Learners]” (p. 238), the use of value-added models to evaluate teachers of emergent bilingual students is likely unreliable and invalid.

However, regarding emergent bilingual students a more consistent source of test invalidity is language bias within the assessments themselves. For example, Menken’s (2010) word frequency analysis of the New York statewide Regents exam found the exam was not only a test of content knowledge but also of English language skills that, by definition, emergent bilingual students are still developing. Abedi (2004) reached a similar conclusion through a different approach, using descriptive tracking of scores and internal consistency analyses of
Annual Yearly Progress as mandated by No Child Left Behind, finding that because of language factors test results were not directly comparable between emergent bilinguals and their peers. Similarly, in a correlational study of the effect of language demands and SAT math scores, Tsang, Katz, and Stack (2008) found that even bilingual students who achieved above the national average in math were still disadvantaged by “language interference” (p. 19) in math word problems, indicating that districts and schools may be labeled as failing due to the size of their emergent bilingual population and the test biases they face (Fairbairn & Fox, 2009). Even accountability models that prioritize growth scores rather than stand-alone outcomes on single tests are likely disadvantaging emergent bilingual students according to Lakin and Young (2013), whose quantitative comparison of targets used by different growth models used in California found evidence that growth models might not accurately reflect future projections of student achievement for emergent bilingual students, thus subjecting these students to unrealistic growth targets that are much higher than their non-bilingual peers.

In addition to issues of implementation and test validity, research has shown that accountability ratings themselves might not lead to the sort of public access to reliable information about school quality that was one of the goals of the accountability movement. The format of the accountability rating (e.g., whether the rating is presented as a letter grade or a proficiency score) affects public interpretation of school quality, as shown through an ANOVA analysis of results from a population based survey of 59 school report cards (Jacobsen, Snyder & Saultz, 2014). Beyond the lack of consistent interpretability of accountability ratings, what the ratings themselves reflect may not even be valuable to public audiences in the first place, as the majority of the public who participated in a mixed methods survey reported they do not see
standardized test scores as indicators of school quality, with less than a quarter of respondents endorsing the opposing view (Brewer, Knoeppel & Lindle, 2015).

The divorce between which metrics the public believes indicate school quality and which metrics are used in accountability frameworks reflects the ideological underpinnings of the accountability movement, which historically has positioned the public and especially historically marginalized communities as expendable rather than integral to defining the principles of accountability and its goals (Lipman, 2013). Such findings cast doubt on the idea that accountability policies actually fulfill their markets-based rationale that desirable and undesirable reputations will drive improved performance. In fact, "[m]isinformation about school quality may steer families away from areas they otherwise would have selected" (Glynn & Waldeck, 2013; p. 476) and vice versa, leading to families taking decisions ostensibly based on objective indicators of school quality that in reality can reflect as little as formatting choices (Jacobsen, Snyder & Saultz, 2014).

Even ignoring these concerns about test validity – as does the majority of accountability research on disadvantaged schools according to Huilla’s (2020) content analysis – research on the ability of accountability frameworks to impact learning improvements is mixed. While the implementation of high-stakes accountability structures has been found to have some positive impacts on student achievement, whether or not this is due to the accountability frameworks themselves or other contextual factors remains undetermined, an ambiguity heightened by the fact that, internationally, differences in quality of instruction only account for approximately 10% of inter-school test score variability (Wiliam, 2010).

Although some research has found that accountability is shown to increase achievement on tests, such as demonstrated by the 2001 study by Fuller and Johnson that analyzed high stakes
test outcomes, advanced placement course patterns, and college entrance examinations, it is not clear that even these findings of improvement “imply that all accountability systems will drive improvement in student achievement. They will not” (Fuller & Johnson, 2001; p. 281). Instead, each accountability system must be weighed for its potential benefits as well as harms, since even accountability systems which are shown to increase performance can have negative consequences. This dynamic was highlighted by Hanushek and Raymond (2005), whose study estimating accountability effects using data from the National Assessment of Education Progress found that – although in the aggregate accountability led to more achievement growth than would have occurred without it – it also led to widening racial achievement gaps as Black and Latinx students gained less from accountability than their White peers. These findings reflect those of Lee and Wong (2004), whose analysis of state policy surveys and achievement data showed that accountability policy had no effect on reducing the achievement gap.

**Research Regarding the Efficacy and Outcomes Of Accountability Policies for Historically Marginalized Students**

Because this dissertation focuses on the impact of local accountability policies on historically marginalized populations with a special attention to emergent bilinguals, it is important to also understand research specifically for these students. Unfortunately, when accountability efficacy and outcomes are considered for these students in particular, research has come to even more dire conclusions, including that accountability frameworks can reduce rather than enhance access to high quality education for historically marginalized students like emergent bilinguals and students with special needs. By definition, these students’ unique needs defy the logic of standardization, yet the standardization which is the bedrock of accountability is continually applied to assessing them, their teachers, and their schools, often with inaccurate
results that are detrimental to the students themselves (Cramer, Little & McHatton, 2018). Overlooking specific populations’ different needs leads to accountability practices which are counterproductive or even harmful, such as by dissuading local leaders from implementing culturally and linguistically relevant interventions that are more likely to be successful, as found in the case study of a successful turnaround of a ‘failing’ school (Reyes & Garcia, 2014).

Accountability does not only impede culturally and linguistically appropriate interventions but also appropriate instruction. As a result of accountability pressures for students to perform well on English-language exams, in her qualitative study of 10 New York high schools Menken (2006) found that the majority of schools began ‘teaching to the test,’ or winnowing instruction to only those skills and knowledge that would be tested. For example, reducing bilingual education options and replacing English as a Second Language curriculum designed to develop the communicative needs of emergent bilingual students with one based on the English Language Arts curriculum designed for native English speakers. This phenomenon is not limited to emergent bilingual students, as Diamond and Spillane (2004) found in their qualitative study that such ‘teaching to the test’ was more common in low-rated schools, which tend to have higher than average proportions of historically marginalized students. This might be because teachers who work in contexts in which accountability frameworks are implemented especially punitively – with consequences for low test performance including school closure, intervention, and staff turnover – have reported that their opportunities for professional development have likewise been winnowed in order to focus on instruction that will produce higher test results. The consequence of this is that the students in these contexts likewise have fewer opportunities to learn from teachers whose professional abilities are being fully developed, as found by a case study conducted by Jacobs, Burns, and Yendol-Hoppey (2015).
In addition to responding to accountability pressures by ‘teaching to the test,’ some schools and teachers have responded by removing students who they perceive likely to have low test scores. Preventing students from taking high stakes tests is one way schools can “game” accountability. For example, using the tests of statistical significance to prove "discriminatory impact" according to the legal standard, Haney’s (2000) review of Texas’s remarkable outcomes for students found that they were reflections of an increase in low-achieving students being removed from the pool of testing takers through higher retention rates of Black and Latinx students, higher rates of drop outs, and excluding students from testing. In a qualitative study, Vasquez Heilig, Young, and Williams (2012) found that more than two-third of teachers and administrators confirmed that their schools eliminated students who were perceived to lower accountability scores by failing students, implementing ‘waivers’ to excuse low performing students from tests, and encouraging low-performing students to drop out of school, even though threats to report students to US immigration due to perceived undocumented status.

These findings mirror what Vasquez Heilig and Darling-Hammond (2008) encountered in their longitudinal mixed methods study of 25,000 students over seven years, which showed that test scores increased in response to student disappearances as low-achieving students were retained or encouraged to leave school, resulting in Black, Latinx, and emergent bilingual students having the lowest graduation rates. The motivation to push out students in order to “game” accountability outcomes might explain discrepancies in graduation rates per student racial group, as Fitzgerald, Gordon, Canty, Stitt, Onwuegbuzie, and Frels (2013) found using nonparametric repeated measures analysis of variance of three years of data representing between 500-600 schools each year, in which evidence emerged that in schools with large enrollment sizes White students had statistically significantly higher high school completion
rates than Latinx and Black students, differences that represented large effect sizes and that indicated students of color were at particular disadvantages in those school settings. Together, these data indicate that accountability policies create disincentives for schools to work with historically marginalized populations (Menken, 2010).

Because accountability policies have led to treatment of historically marginalized students in which their curriculum and instruction are narrowed, they are erroneously retained or discouraged from participating in school, and they are perceived as threats by teachers and administrators, it is little surprise that accountability outcomes often reflect student populations. In a regression analysis of the relationship between students’ reported Annual Yearly Progress and student demographics, Martin (2012) found that the schools that failed to achieve their benchmarks had larger populations of economically disadvantaged students. Similarly, Harris (2007) used descriptive statistics of student demographics and school ratings of 60,000 schools across 47 states and found that schools with the smallest populations of students of color and students in poverty were 89 times more likely to be rated as high performing when compared to schools with larger populations of these students, indicating that the accountability framework was holding schools accountable for student demographics, a factor beyond their control. These findings were confirmed by Martinez-Garcia, LaPrairie, and Slate (2011), whose MANOVA analysis of student demographics and accountability ratings of 4,000 schools found that, because “[e]xemplary elementary schools had the lowest percentages of Black students, Hispanic students, economically disadvantaged students, at-risk students, students with LEP, and mobility percent whereas Academically Unacceptable schools had the highest percentages” (p. 16) with moderate to large effect sizes, accountability ratings were likely unreliable as they also were reflections of student demographics. Regarding emergent bilingual students, the correlational
study conducted by Tsang, Katz, and Stack (2008) confirms that due to “language interference” (p. 19) the standardized tests used to calculate accountability ratings are unreliable, as these ratings reflect schools’ emergent bilingual populations rather than student learning.

Such disparities in outcomes along the lines of student demographics reflects disparities of inputs. Using a fixed effects regression analysis to evaluate the relationship between student demographics, school resources, and outcomes over ten years, Wu (2013) found as small as a 1% change in student racial groups, students receiving Free and Reduced Lunch, or English learner students was enough to change schools’ accountability outcomes. Unsurprisingly, a similar relationship was detected between schools’ resources, with achievement increasing in proportion with an increase in teachers with full teaching credentials and decreasing with an increase in school enrollment and class size. Together, these findings indicate that accountability ratings are not only a reflection of student learning but also student demographics and school resources.

Quantcrit Approaches To Understanding School Environments and Outcomes for Historically Marginalized Students

The theoretical and methodological framework employed in this dissertation was QuantCrit, which has roots in Critical Race Theory and seeks to use quantitative data for antiracist purposes in part through the acknowledgment that, like all data, quantitative data are not objective but require specific interpretations and reflect researcher intention (Gillborn, Warmington & Demack, 2018). In statistical demographic research, this translates to an attention to how racial categories are understood to be causally related to disparate social outcomes. QuantCrit scholars hold that quantitative data do not “speak from themselves” but, like all data, only become meaningful through interpretation. For this reason, the ways that quantitative researchers discuss and interpret findings – such as the relationship between racial groups and
social outcomes – has powerful implications for either perpetuating white supremacist understandings of inherent racial difference and deficit or disrupting those narratives (Covarrubias, Nava, Lara, Burciaga & Solórzano, 2019). For example, instead of interpreting data as a person’s race “causing” disparate social outcomes, we must insist on interpretations that explore how social processes of racialization are related to outcomes (Bonilla-Silva & Zuberi, 2008). Therefore, a better way to report on statistical findings is through a commitment to coupling discussions about race with discussions about racism (Gillborn, Warmington & Demack, 2018).

This is embodied in recent research into disparities of educational outcomes per student demographics, which Van Dusen, Nissen, Talbot, Huvard, and Shultz (2022) framed as reflections of education debt rather than student ability. Using hierarchical linear models of pre- and post-tests of over 4,000 college students taking introductory chemistry courses across 12 institutions, they found that Black men and women were owed the largest education debts by society. For example, White Hispanic women would need to take the introductory course two and a half times to be repaid their legacy education debt. Another way demographic data can be used within QuantCrit frameworks is to calculate relative difference composition indexes, Equity Indexes, and Inequity Scores as demonstrated by Young and Young (2022), who found Black students were nationally underrepresented in Gifted and Talented programs by between 31%-56%. Similarly, in a 2021 study of math and English Language Arts achievement for different demographic groups, Ramlackhan and Wang (2021) used descriptive statistics and growth mixture models with varying numbers of latent classes to find that student demographics varied dramatically across the higher- and lower-achievement classes, with the high-achievement classes being overrepresented with White students. However, instead of attributing these
differences to the students themselves, the authors used a QuantCrit framework to call for greater investigation of the “underlying structures and oppressive mechanisms in society that creates differential access to resources and opportunity in urban communities” (Ramlackhan & Wang, 2021; p. 22).

Other recent research has drawn on the QuantCrit view that neither quantitative data nor racial categorizations are neutral or objective to explore the process of student racialization in educational reporting on student demographics. For example, Campbell-Montalvo (2020) used a QuantCrit lens to find that actors across schools and districts interpreted and reported on student race differently, resulting in inconsistencies that revealed the artificial nature of racial demographics in school data and consequently in race-based educational policy. Likewise, a study by Crawford (2019) examined the deeply political and biased nature of educational statistics in her investigation into much-publicized data showing “discrimination” against White, working-class, male students in the UK. This study found that the statistics had failed to disaggregate for status of educational attainment while misrepresenting class status in order to present White males as if they were chronically underserved in public schools when in reality they continued to outperform their peers of color, making the statistics little more than misleading data serving to legitimize the centering of White needs through a false sense of White, male victimhood.

Additionally, many QuantCrit scholars borrow from theories of intersectionality, which holds that our various raced, classed, gendered, sexual, linguistic, religious, etc. social identities intersect in different arenas of our daily lives to create overlapping and at times contradictory spheres of oppression and privilege (Crenshaw, 1991). Demographic statistics can fail to account for intersectionality, such as through the misuse of multiple regression analyses that seek to
isolate and control for identity variables as if they were discrete and extractable aspects of social existence rather than necessarily interconnected (López, Erwin, Binder & Chavez, 2018). QuantCrit scholars have attempted to address this shortcoming through quantitative intersectionality, a method and approach that is grounded in intersectionality through the purposeful examination of how various social identities lead to differential outcomes rather than treating singular identities, such as race or gender, as if there were homogeneous social categories that relate to homogeneous experiences (Covarrubias, Nava, Lara, Burciaga & Solórzano, 2019).

Using an intersectionality framework, several scholars have problematized the misleading homogenization of students according to gender, race, class, and immigration status to identify statistical differences in education attainment trends according to intersectional identities. López, Erwin, Binder, and Chavez (2018) used saturated logistic models in an analysis of a large public university’s six-year graduation rates and developmental coursework enrollment, finding that students’ intersectional identity categories greatly related to their likelihood of graduating, with low-income American Indian men being approximately 45% less likely than high-income White women to graduate in six years.

An especially fruitful area for quantitative intersectionality focuses on the educational pipeline in which percentages of academic attainment can be tracked to intersectional identities that holistically analyze the social locations of, for example, female, upper-income, noncitizen, Chicanas as opposed to male, middle-income, citizen, African Americans (Covarrubias, 2011). Framing quantitative investigations in this way not only offers to produce a richer, more accurate understanding of how identity categories relate to differential outcomes, but it also disrupts majoritarian narratives that essentialize historically marginalized communities to singular aspects.
of their identities, a disruption that is a central goal of Critical Race Theory and QuantCrit (DeCuir & Dixson, 2004; Garcia, López & Vélez, 2018). Other research of the education pipeline has used similar methods to find that gender, class, and citizenship status directly relate to educational attainment and earning potential, whether for Asian Americans (Covarrubias & Liou, 2014), students of Mexican origin (Covarrubias & Lara, 2014), or across racial groups (Covarrubias, 2011) throughout the K-PhD spectrum.

These lenses also lend themselves to critical studies of proportionality. Cruz, Kulkarni, and Firestone (2021) used mixed multilevel logistic regression models and discrete-time hazard models to find that BIPOC students were overrepresented in both in- and out-of-school suspensions, representing a form of instructional loss for these students. Even when controlling for gender, Free and Reduced Lunch status, parent education, and school characteristics such as the percentage of White students and average years of teacher experience, they found that disciplinary actions tracked student race, with Black and Latinx students being about twice as likely to be suspended as White students, with the odds of any student being suspended decreasing in proportion to increases in the percentage of the student body that was White. These results mirrored those of Anyon, Wiley, Samimi, and Trujillo (2021), who also used descriptive statistics and multilevel logistic regressions to calculate odds ratios of being suspended per student demographic, finding that when compared to White students, Black, Latinx, and multiracial students had significantly higher odds of receiving both in- and out-of-school suspensions. Both of these studies were grounded in QuantCrit, and thus interpreted these discrepancies not as inherent attributes of students but reflections of systemic inequities within public education policies and practices.
Other work highlights the inaccuracies of attributing disparities to students, such as the multilevel multivariate logistic regression study conducted by Morris (2021), which found that, according to the results of the nationally representative Education Longitudinal Study, the belief that students of color were more likely to learn in disruptive, violent schools necessitating these disparate discipline rates fell apart, as "students who attend minority segregated schools are, at worst, no more likely to be victimized, and, once statistical controls are put into place, they appear less likely to be victimized” (p. 13) by other students at their schools. This work quantitatively demonstrates the fallacy of attributing disparate disciplinary outcomes to students themselves, an example of the power of QuantCrit to both highlight educational inequities while directly contesting deficit narratives. The power of such work is not limited to students but can also be applied to all participants in the educational system, as demonstrated by Campbell-Montalvo (2020), who used hierarchical linear modeling to describe disparities in teacher evaluations as found in North Carolina Department of Public Instruction administrative data, finding that even when quality of teaching indicators are similar, after classroom observations are conducted Black women are rated lower than White women. Campbell-Montalvo draws on QuantCrit to interpret these data not as implications of inherent racial difference but rather inherent racial bias within public school settings that affect all participants, students and teachers alike.

**Relationship Between Previous Research and the Dissertation**

This dissertation drew from QuantCrit in its design, interpretation, and purposes. Because this project explicitly aimed to understand the ways that the School Performance Framework accountability system is measuring student demographics rather than student learning, its primary purpose is to advance critical scholarship that problematizes school accountability
policies and frameworks as metrics of social and institutional biases that legitimate the maldistribution of educational resources both symbolic (e.g., high accountability ratings and prestige) and material (e.g., appropriate quality curriculum) and perhaps even the cessation of programs that have potential to be effective for students such as bilingual or dual language.

In this way, this research was fundamentally dedicated to promoting social justice causes as it sought to explore the ways that institutional practices in the form of school accountability frameworks result in disparate outcomes for historically marginalized populations while benefiting historically dominant populations. By highlighting the systemic nature of institutional bias in this context, this research project offers empirical evidence that counters deficit narratives about historically marginalized communities and points to policy reforms that account for and thus overcome such biases.

Additionally, this project looked at all available data regarding student demographics, including racial designations, class designations (as approximated by Free and Reduced Lunch status), language designations (as measured by “English Learner” status and WIDA scores in English proficiency and Spanish-language status), and ability designations (as measured by Special Education status), as well as intersectional identities, such as the percentage of English Learners who receive Special Education services.

Doing so grounded this project in an intersectional lens in which student identities are not seen as discrete but rather interwoven constellations of social locations that combine to impact educational outcomes. Such a focus not only allowed this research to produce more robust descriptions of patterns of institutional bias against raced, classed, and linguistically- and ability-marked students, but it also embodied a priority to disrupting essentializing portrayals of historically marginalized populations in research. Both of these goals and the intentional research
design and data collection procedures they inspire are derived from scholarship on intersectionality generally and in QuantCrit specifically.

Finally, this project drew on CRT and QuantCrit regarding the need for scholars to take explicitly political, antiracist stances during all stages of the research process, including during the analysis and interpretation phases. For this reason, during analysis and interpretation this project only looked for and discussed relationships between historically marginalized identity categories and disparate academic outcomes in terms of racism and racialization rather than racial causation. This commitment was extended to other social processes of marginalization such as classing and linguistic discrimination as warranted by the data. At no time did this dissertation entertain the possibility that students’ racial, class, language, or ability statuses “cause” disparate accountability outcomes. Rather, any relationship found between those statuses and accountability outcomes was investigated as reflections of institutional biases against such populations that relate to the lack of construct validity in accountability frameworks, inadequate supports for schools that serve large numbers of such identified students, or both.
Methods

Data Overview and Study Parameters

To produce antiracist, social justice scholarship, to acknowledge the complexity of both student social identities and school contexts, as well as the historical context of Denver Public Schools’ struggle to equitably serve historically marginalized students, this study used a transformative research design (Teddlie & Tashakkori, 2009) based in QuantCrit principles (Gillborn, Warmington, & 2018). Such quantitative data analysis allows for a better understanding of the ways the accountability framework used by DPS inadvertently measures non-academic variables such as student demographics, English Learner (EL) characteristics and services, and school contexts instead of student learning. Doing so allowed this project to highlight the ways that the School Performance Framework (SPF) is a measure of variables extrinsic to the accountability framework, policy, and purposes, thus rendering the SPF a reflection of institutional biases that reproduce inequality in education rather than student strengths and needs.

Research Questions

In order to investigate the relationship between accountability outcomes and student demographics, EL characteristics and services, school contexts, this study addressed the following research questions:

1. What are the student demographics, EL characteristics and services, and school contexts per SPF rating bracket?

2. At what rate do schools remain in, enter into, or exit the most extreme SPF ratings statuses of Intervention vs. Blue, and what are the student demographics, EL characteristics and services, and school contexts in them?
3. What are the student demographics, EL characteristics and services, and school contexts per charters and district-run schools?

4. Do student demographics predict SPF scores?

5. Do student demographics predict SPF outcomes?

Timeframe

This study drew on data from the three most recent academic years available during which the School Performance Framework (SPF) was implemented in Denver Public Schools consistently. They are the 2016-2017, 2017-2018, and the 2018-2019 academic years (AYs). After the 2018-2019 AY, COVID resulted in disruptions that made many of the metrics used in the SPF unreliable, resulting in no SPF scores being issued for the 2019-2020 AY (Denver Public schools, n.d. - d) Before the 2016-2017 academic year, the district implemented several changes to how accountability scores were calculated that made comparison across years problematic. These changes included (a) the addition of the new Equity Indicator, (b) switching from using the Partnership for Assessment of Readiness for College and Careers (PARCC) standardized tests to the Colorado Measures of Academic Success (CMAS) standardized tests to calculate SPF scores, and (c) lowering the threshold as to what constitutes adequate performance on some measures (Asmar, 2016). Because of the changes in how accountability scores were calculated in the prior years and the global disruptions to education in the subsequent years, the span of academic years 2016-2017 through 2018-2019 represent the most recent years in which accountability policy has been consistently applied in the district.

In this study, each individual academic year is represented with an individual dataset. In the study’s use of descriptive statistics, individual academic year’s data trends are shown, as well as the averages derived when the three years of data are aggregated. In the use of regression
models, the three-year aggregate is used, and dichotomous variables are included to control for different years.

Data Sources

Only publicly available school-level data pertaining to the district were included. Data were drawn from various datasets across three sources: (a) the Colorado Department of Education (CDE) publicly available online data of school-level staff, discipline, and student statistics; (b) DPS annual SPF Reports; and (c) Consent Decree reports of “English Learner” services and outcomes as mandated by the Modified Consent Decree (2012) related to mandated services, programs and assessments for students identified as English Learners (Consent Decree of the U.S. District Court (2012). Data represent both district-run schools and charters. In total, nine datasets were used to compile each academic year’s final dataset: four datasets came from the CDE, one dataset came from the SPF Report, and four datasets came from the reports mandated by the Consent Decree. A summary of data sources can be found in Appendix Table 1 (Appendix A).

Inclusion Criteria

Because this dissertation examined the relationship between accountability outcomes in the form of SPF scores and student demographics, EL characteristics, and school contexts, the principle inclusion criteria was the availability of SPF accountability scores. Due to the diversity of reporting sources and datasets, when all nine datasets were combined into final combined datasets for each academic year, these were consistently incomplete. For example, although a school might have had data from the SPF report, the CDE datasets, and most of the Consent Decree datasets, perhaps in the dataset of English Learner participation rates in Gifted and Talented programs there could have been no entry for that school. In this case, the school still
would have been included, as the secondary inclusion criteria was data available in at least one additional dataset beyond the SPF Report. Fortunately, all schools with SPF scores met this secondary criteria.

Exclusion criteria

As only schools with reported SPF scores were included, any school lacking this data was omitted. This resulted in the omission of 30 schools in the 2016-2017 AY, 17 schools in 2017-2018 AY, and 24 schools in the 2018-2019 AY. In addition, in some datasets multi-level schools (e.g., serving grades K-8) were reported as a single entity while in others the levels were disaggregated and reported separately. For example, in the SPF Report in the 2018-2019 AY there is an entry for “Bruce Randolph School,” but in the same year the 9VA2 Consent Decree report lists “Bruce Randolph HS” and “Bruce Randolph MS.” Because there was no way to discern which outcomes of the two or more entries in the disaggregated reporting would be relevant to which outcomes in the single, aggregated reporting, when there was reporting inconsistency of multilevel schools those schools were omitted. This resulted in the omission of 38 schools in the 2016-2017 AY, 13 schools in 2017-2018 AY, and 14 schools in the 2018-2019 AY.

Variable Type and Selection

These datasets were chosen because they provided the variables necessary to address the research questions of the study. The variables used can be categorized broadly into the following themes: (a) student demographics, (b) English Learner characteristics and services, (c) school contexts, and (d) accountability outcomes.
Student Demographics

Student demographic variables were defined by the respective reporting agency (i.e., the CDE or DPS) and included counts of students classified as Students of Color (SoC), English Learners (EL), Special Education (SPED), Gifted and Talented (GT), and Free and Reduced Lunch (FRL). These variables were included for several reasons. First, research has shown that students in the SoC, EL, FRL, and SPED classifications are historically denied the equitable services, opportunities, and resources necessary for school success (Darling-Hammond, 2004; Martin, 2012; Wu, 2013), with disparate opportunities to participate in GT programming being an additional hallmark of this inequitable allocation of resources (Card & Giuliano, 2016) and a rationale for including the GT metric in the study. Second, because this study pays special attention to ELs, including variables regarding SoC, FRL, and SPED are relevant due to frequent intersectional categorizations in which ELs occupy one or more of these additional classifications (Blanchett, Klingner & Harry, 2009; Cramer, Little & McHatton, 2018). Finally, the outcomes of students of color, students receiving FRL services, ELs, and students receiving SPED services are specific indicators used by DPS to calculate an important part of the SPF score called the Equity Indicator (Asmar, 2016b). Because of the centrality of these student demographic classifications in research on equitable education, their importance to ELs specifically, and their role in calculating SPF scores, all of these variables were included in the study as “Student Demographics.” The raw counts of these student demographics were transformed into percentages of each student demographic type out of the total student population. In this study, all four of these classifications of historically marginalized students (SoC, ELs, FRL, SPED) are used as predictors of SPF scores in multiple regressions in addition to reporting of descriptive statistics.
EL Characteristics and Services

This dissertation also drew on the expansive accountability reporting mandated by the Modified Consent Decree of 2012 (Consent Decree of the U.S. District Court, 2012) regarding the characteristics, needs, and outcomes of ELs. Since ELs are typically overrepresented in SPED programming (Sullivan, 2011) and underrepresented in GT programming (US Commission on Civil Rights, 2018), this study included metrics describing EL participation rates in these programs. The district reported the percent of ELs classified as GT and the percent of SPED students that were classified as ELs in each school. In addition, this study reports on the language status of ELs to specify when their bilingualism includes Spanish, as raciolinguistic ideologies that index racial status by language practice has led to English-Spanish bilingualism being especially denigrated in the US when embodied by heritage speakers of color (Hill, 2009; Rosa & Flores, 2017).

This study also includes variables describing Parent Preference 1, 2, and 3 (PPF1, PPF2, PPF3), which indicate what kinds of language supports parents desire for their EL students, with PPF1 indicating a preference for native language instruction designed for emergent bilingual students, PPF2 indicating a preference for English-only instruction designed for emergent bilingual students, and PPF3 indicating a desire to decline all services offered specifically to ELs. These data are paired with EL participation rates in the settings of what the district calls Mainstream (reflecting PPF3), English Language Acquisition-English (reflecting PPF2), English Language Acquisition-Spanish (reflecting PPF1), and Dual Language programs. Access to language instruction settings is particularly important to ELs, who are often denied opportunities to participate in challenging and appropriate curriculum (Callahan & Hopkins, 2017).
Additionally, data describing the rates at which ELs were Redesignated from, Exited from, and Re-Entered into EL status were also included, as these rates can reflect policy and instruction that impacts both ELs’ opportunities to access challenging curriculum as well as their achievement outcomes (Brooks, 2020; Kim, 2017). Finally, EL data describing WIDA ACCESS scores – which measure English-language proficiency across the domains of reading, writing, speaking, and listening – were also included to indicate the percentage of ELs that were Beginning, Intermediate, and Advanced Level in their development of English, as prior research has shown these distinctions to be statistically significant predictors of SPF outcomes (Strong & Escamilla, 2020). In this study, all of these variables were calculated as percentages reflecting rates out of the total EL school population in each school and used in descriptive statistics to illustrate differences in EL characteristics and services across schools.

**School Contexts**

In order to describe the variation across school settings as well as to provide controls for multiple regression models, this study also included variables regarding school characteristics. Because the race and socioeconomic status of students has been found to predict rates of disciplinary referrals (Bryan, Day-Vines, Griffin & Moore-Thomas, 2012; Skiba, Chung, Trachok, Baker, Sheya & Hughes, 2014) one such characteristic describes disciplinary actions and incidents to represent whether students are learning in particularly discipline-heavy environments, which might be related to bias against students of color and students in poverty. All disciplinary action and incident counts were converted into rates of actions and incidents per 100 students. The discipline action counts were also used to calculate a new variable to describe the rate of disciplinary actions that resulted in instructional loss per 100 students, since some types of discipline such as out of school suspensions result in considerable loss of access to
teachers and instruction, making the disparate rates of discipline students of color confront equivalent to the loss of months or even more than a year of instructional time (Losen & Martinez, 2020). This variable was created to capture an additional potential impediment to learning outcomes that could influence SPF ratings, a factor that individual analysis of disciplinary action counts in isolation would obscure. This variable was made by combining the counts of disciplinary actions of expulsion, out of school suspension, and classroom removal, and then calculating the rate of those aggregated counts per 100 students. All discipline variables were included in descriptive statistics, and the variable describing loss of instructional time was also included as a control in the multiple regressions.

Another variable included in the descriptive statistics describes total school enrollment. This was included both because of previous research that has found that enrollment size can impact relative disadvantages of students of color (Fitzgerald, Gordon, Canty, Stitt, Onwuegbuzie & Frels, 2013), and because the district is currently considering closing schools with small enrollment (Asmar, 2021), making it especially pertinent to immediate district interests and considerations when defining school success. Additionally, this study uses variables describing student-teacher ratios and the percentage of teachers that are considered “Fully Qualified” to work with culturally and linguistically diverse students according to the district. The teacher qualification metric was selected because students in poverty, students of color, and emergent bilingual students are less likely to work with highly qualified teachers (Darling-Hammond, 2004; Goldhaber, Lavery, & Theobald, 2015; Lankford, Loeb & Wyckoff, 2002). The student-teacher ratio metric was included due to previous research that found that these ratios are related to student achievement and teacher stress (Alspaugh, 1994; Hojo, 2021; Koc & Celik, 2015). Finally, this study also included a dichotomous variable to describe whether or not
a school was district-run or a charter in order to address Research Question 3. While all of these variables were included in descriptive statistics, only the percent of teachers classified as “fully qualified,” the student-teacher ratio, and the rate of disciplinary actions that result in instructional loss per 100 students were also included in the multiple regressions as controls. The theoretical and data-based decision making process regarding model construction will be discussed in the section describing methods for each research question.

Variables Created

As mentioned, most of these variables were transformed from counts into rates and percentages in order to standardize occurrences across schools of different sizes, although some new variables were also created. For example, a new variable was created to represent Simplified SPF Ratings designations by (a) collapsing the ratings categories (Red and Orange) that result in district intervention into one category, called “Intervention;” (b) leaving the middle rating category (Yellow) as a single category, called “On Watch;” and (c) collapsing the two highest categories (Green and Blue) into one category, called “High Performing.” These were created both because there are not meaningful differences between the collapsed ratings categories as they result in similar outcomes (such as prestige, stigma, or intervention), as well as to run ordinal logit regressions, which predict categorical outcomes, with results that were easier to interpret as they showed predicted probabilities of broadly distinct accountability outcomes rather than framing ratings outcomes with similar results as somehow different.

In a similar way, a set of variables was created to describe SPF trends over time. To do so, I looked at schools that remained in, entered into, or exited from the SPF ratings categories at the most extreme ends of high and low performance. At one extreme, schools were coded to describe if they remained in, entered into, or exited Intervention Status (as defined above), and at
the other they were coded to describe whether they remained in, entered into, or exited Blue Status, the most exclusive and thus most prestigious designation at the other pole of accountability outcomes. The rationale and use of these categories will be discussed below in the section for each research question.

Research Process

Data Collection and Cleaning

Each academic year’s nine datasets (27 in total) were downloaded from the three data sources as Excel files. If rows did not contain individual school cases or contained nested data, the Pivot tool of Excel was used to clean the data so that each row only described single school cases. Values of “0” were inspected to ensure they actually represented a count or percent of 0 and not the absence of data. In the few cases in which values of “0” represented an absence of data, the value was deleted and a blank space was left in its place.

Because of inconsistency in how school names were reported across the nine datasets, I used Excel to clean the name text for each school. First, I used the UPPER function to capitalize all school names. Then, I used the replace tool to ensure all instances of school level descriptions were consistent, as in some datasets a school could be described as, for example, “Lincoln High School” and in other datasets as “Lincoln HS.” Finally, I used the TRIM function to remove additional spaces. This resulted in consistent reporting of school names across the datasets.

Data Consolidating (Excel)

These cleaning procedures allowed me to use the Consolidate tool of Excel to combine all nine datasets for each academic year, using the school name as the identifying metric. Although a preferable case identifier would be a numeric code, for reasons I do not have access
to DPS uses school level numeric identifiers that are different than those used by the Colorado Department of Education, resulting in two sets of irreconcilable identifiers that only a name-by-name check could match, a process which would have introduced unacceptable degrees of human error.

Data Merging (Stata)

Once there was a single, complete dataset for each academic year, I imported them into Stata by holding the three datasets in memory and creating dichotomous variables to indicate each distinct academic year. This allowed me to calculate averages, run regressions, and conduct analyses for the three-year aggregate as well as conduct analyses and output for individual years.

Stata Functions

I then used Stata to create variables to transform data from counts into percentages and rates of student demographics and service types, discipline per 100 students, and fully qualified teachers. Some student demographic variables, like percentages of Spanish-speaking ELs and ELs receiving Special Education services, reflect the percentages of these students out of the total number of their respective subpopulations (i.e., ELs) rather than the total number of students enrolled.

Stata was also used to create the new variables, like SPF Simplified Outcomes and SPF trends. Some of these new variables required several calculations, such as the variables describing the percentages of ELs according to level of English proficiency, which were created by combining the counts of ACCESS scores of: (a) 1 and 2 to create the count of Beginning Level ELs, (b) 3 and 4 to create the count of Intermediate Level ELs, and (c) 5 and 6 to create the count of Advanced Level ELs. These counts were transformed into percentages to represent the rates of ELs in each level of English proficiency out of the total number of ELs in a school.
Stata was then used to create tables of descriptive statistics in order to address Research Questions 1, 2 and 3, and run the multiple regressions required to address Research Questions 4, 5, and 6. Stata was also used to export the data used in all of the tables and the figures created in R Studio (below), and to create the figures used in Research Questions 4 and 6.

Creation of Figures

R Studio was used to create figures for the descriptive statistics in Research Question 1 and the predicted probabilities resulting from the ordinal logit regressions in Research Question 5.

Methods per Research Question (RQ)

RQ1. What are the student demographics, EL characteristics and services, and school contexts per SPF rating bracket?

To address this research question, I used Stata to create descriptive statistics of the mean of the variables described in the above section per each of the five SPF ratings brackets (Red, Orange, Yellow, Green, Blue). Results were exported to Excel and reported in a table in order to show each individual academic year’s means as well as the three-year aggregate means. These data were then used to create figures in R Studio using ggplots.

RQ2. At what rate do schools remain in, enter into, or exit the most extreme SPF ratings statuses of Intervention vs. Blue, and what are the student demographics, EL characteristics and services, and school contexts in them?

To address this research question, schools were coded to describe whether they remained in, entered into, or exited either Intervention Status or Blue Status. These two statuses were chosen to represent the poles of accountability outcomes as means of evaluating the effectiveness
of the SPF accountability framework in discouraging school failure and promoting school success (Murray & Howe, 2017). At one pole is “Intervention Status,” representing the schools receiving either Red or Orange SPF ratings, which trigger district intervention (Asmar, 2018; Denver Public Schools, 2018). As such, schools in the Intervention Status category represent primary targets of the accountability framework; namely, if the accountability framework is effective in promoting school success, schools should receive Intervention Status only temporarily as the accountability consequences of low ratings promote higher levels of success. At the other extreme are the schools that earned the highest rating possible, or “Blue Status,” and thus were used to represent the end toward which the accountability framework should, in theory, move schools and at which schools should aspire to remain. Together, Intervention Status and Blue Status not only represent the extremes of the SPF accountability system but also the targets of that system.

To conduct this analysis, I created a variable “SPF Trends” and gave a non-ordinal numeric code to schools that either (a) remained in Intervention Status, (b) remained in Blue Status, (c) entered into Intervention Status, (d) entered into Blue Status, (e) exited Intervention Status, or (f) exited Blue Status. Schools were coded as “remaining” in either status if they began and ended the study timeframe in that same respective status. They were coded as “exiting” one of those statuses if they began the study in that status and ended with any other SPF rating. They were coded as “entering” one of those statuses if they began the study in a different SPF rating and ended the study in either Intervention or Blue Status.

Schools that did not meet any of these criteria were coded as 0. Only schools with SPF data for all three years of the study were eligible to receive non-zero numeric codes, as the research question seeks to identify trends over time and even missing a single year’s data would
result in trends only describing year-over-year change, which I decided was not sufficient to constitute a trend rather than potential noise.

I then used Stata to create and export descriptive statistics of each SPF Trend category for each individual year and the three-year aggregate means. I also used Stata to create quartiles of the variables, which I used to run crosstabs of each of the SPF Trend categories per the quartiles. In doing so, my aim was to triangulate the findings, thus showing that differences in student demographics, EL characteristics and services, and school contexts did not only represent potentially insignificant variation of a few percentage points but indeed reflected schools at the extremes of the distribution of the study variables.

**RQ3. What are the student demographics, EL characteristics and services, and school contexts per charters and district-run schools?**

Similar methods to those used to answer Research Questions 1 and 2 were also employed to address Research Question 3, as all three of these research questions resulted in the creation of descriptive statistics of the means of the study variables for each individual academic year as well as the means for the three-year aggregate. For this research question, I used Stata to create and export the means for all the study variables per two categories: whether a school was district-run schools or a charter. Stata was then used to create and export these data for each individual academic year as well as the three-year averages.

**RQ4. Do student demographics predict SPF scores?**

To address Research Question 4, Stata was used to run OLS multiple regressions to test whether student demographics predicted the percent of SPF points earned. Individual student demographic predictors were the percent of the student population classified as (a) Students of Color, (b) English Learner, (c) Special Education, or (d) Free and Reduced Lunch.
These regressions held constant: (a) the percent of teachers that are classified as Fully Qualified, (b) the student-teacher ratio, (c) the number of disciplinary actions that result in instructional loss per 100 students, and (d) dichotomous variables for the academic years 2017-2018 and 2018-2019, with the variable for 2016-2017 being omitted as the reference. Although there were other school context variables present in the study that could have served as alternative or additional controls, the decision to include or omit some of these variables as controls in the multiple regression models was both based in the data and in theory.

I decided not to include rates of disciplinary actions and incidents as controls because the disciplinary outcomes that could most directly impact learning achievement was already captured through the variable of disciplinary actions resulting in instructional loss. In addition, this latter variable had a high degree of collinearity with the former discipline variables (r=0.77 and r=0.75 respectively) (Table 1), making them problematic additional controls. Similarly, the categorical variable indicating whether a school was a charter or district-run did not have a statistically significant correlation coefficient with the outcome of interest, the percent SPF points earned (Table 1), and there was not a sufficient rationale in the existing research literature to justify its inclusion despite its lack of a statistically significant correlation with the outcome of interest. Finally, the enrollment variable was not included both due its small correlation coefficient (r=0.10), which implies a lack of practical significance despite its statistical significance, in addition to the limited extant research literature regarding its mediating role in accountability outcomes that could justify its inclusion on theoretical grounds. As seen in Table 1, the controls chosen in these regressions all had statistically significant correlations with the percent of SPF points earned with moderate to large coefficients, a data-based rationale for inclusion that complemented the theoretical and research-based rationales regarding how they
could impact achievement outcomes independent of yet related to student demographics, thus representing alternate metrics that could influence the learning outcomes that the accountability policies aim to measure apart from the student demographic predictors.

**Table 1.**

*Pearson Correlations of Potential Control Variables and SPF Percent Points Earned*

<table>
<thead>
<tr>
<th>SPF Points Earned %</th>
<th>SPF Points %</th>
<th>Disc. - Loss</th>
<th>Full Qual. Teachers %</th>
<th>Student-Teacher Ratio</th>
<th>School Type</th>
<th>Enrollment</th>
<th>Disc. - Incidents</th>
<th>Disc. - Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discipline - Instructional Loss per 100 Students</td>
<td>-0.25*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fully Qualified Teachers %</td>
<td>0.18*</td>
<td>-0.25*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student-Teacher Ratio</td>
<td>0.16*</td>
<td>-0.16*</td>
<td>0.17*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School Type (charter or district-run)</td>
<td>-0.08</td>
<td>0.19*</td>
<td>-0.58*</td>
<td>-0.21*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrollment</td>
<td>0.10*</td>
<td>-0.09*</td>
<td>0.15*</td>
<td>0.33*</td>
<td>-0.24*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discipline - Incidents per 100 Students</td>
<td>-0.21*</td>
<td>0.75*</td>
<td>-0.31*</td>
<td>-0.18*</td>
<td>0.18*</td>
<td>-0.05</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Discipline - Actions per 100 Students</td>
<td>-0.23*</td>
<td>0.77*</td>
<td>-0.29*</td>
<td>-0.18*</td>
<td>0.18*</td>
<td>-0.03</td>
<td>0.96*</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* Indicates p-value ≤ 0.05

These regressions used a series of student demographic variables to predict the percentage of SPF points earned when these control variables were held constant. First, I ran two models for each individual student demographic predictor and the controls, one in which the student demographic predictor used linear terms and one in which the student demographic predictor used cubed terms. Cubed terms were chosen due to the apparent nonlinear relationship between each student demographic predictor and the outcome, the percent of SPF points earned, as evident in the spread of the scatterplots showing the relationship between each student demographic predictor and the percent of SPF points earned (Figure 2).
Then, I created two sets of saturated models using all the student demographic predictors together along with the controls. However, due to high collinearity between the Student of Color and the Free and Reduced Lunch variables (r=0.95) as shown in the Pearson’s correlations in Table 2, both could not be included in a single model. This prompted me to create two sets of saturated models: One using the Student of Color variable, and the other using the Free and
Reduced Lunch variable. Like the models of individual student demographic predictors, I first created a model using linear terms for all the student demographic predictors and then squared terms and also cubed terms when they were statistically significant.

Table 2.
*Pearson Correlation of Student Demographic Predictors and SPF Percentage Used in Multiple Regressions*

<table>
<thead>
<tr>
<th>Student Demographic Predictor</th>
<th>SPF %</th>
<th>Student of Color %</th>
<th>Free % Reduced Lunch %</th>
<th>English Learner %</th>
<th>Special Education %</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPF %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student of Color %</td>
<td></td>
<td>-0.37 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free % Reduced Lunch %</td>
<td></td>
<td>-0.38 *</td>
<td>0.95 *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English Learner %</td>
<td></td>
<td>-0.19 *</td>
<td>0.75 *</td>
<td>0.77 *</td>
<td></td>
</tr>
<tr>
<td>Special Education %</td>
<td></td>
<td>-0.40 *</td>
<td>0.28 *</td>
<td>0.34 *</td>
<td>0.07</td>
</tr>
</tbody>
</table>

* Indicates p-value ≤ 0.05

Finally, I used each model of individual student demographic predictors using cubed terms (as opposed to the models using linear terms) to create predicted margins, which then were employed to create figures with Stata to show how changes in the student demographic predictor relates to changes in the predicted SPF percent points earned.

*RQ5. Do student demographics predict SPF outcomes?*

To address Research Question 5, the same student demographic predictor variables and same controls were used to run ordinal logit regressions with the outcome of Simplified SPF categories of (a) Intervention Status, (b) On-Watch Status, and (c) High-Performing Status.

The first step toward addressing this question involved the creation of a new series of Simplified SPF ratings categories. This was done to capture the similarities of accountability outcomes rather than treating those similarities as artificially distinct. For example, a school is
subject to poor repute and district intervention if it receives either a Red or Orange SPF rating. The Simplified SPF ratings categories also reflected the ways the district describes similarities between SPF ratings outcomes. On the DPS website, Blue and Green ratings are described as representing a similar accountability outcome, as they are “the top ratings,” each indicating that a school “is generally doing well in the areas of student academic growth, family satisfaction, equity, and more,” and each representing the points toward which all schools should aspire as “all schools are working to achieve Green or Blue ratings” (Denver Public Schools, n.d. - b). Likewise, DPS describes both Red and Orange rating as indicating the “need of significant improvement” (Denver Public Schools, n.d. - c) because “the school needs a lot of extra support to improve,” which initially comes in the form of an Improvement Plan, and, if that is not successful, then “DPS may also need to make significant changes to the school program or leadership. If a school receives a Red or Orange rating for several years, then DPS may restart or close the school” (Denver Public Schools, n.d. - c). For these reasons, I decided that combining similar SPF ratings outcomes would be the most efficient means of addressing this research question, as it seeks to explore the relationship between student demographic predictors and accountability outcomes and the district describes multiple ratings as resulting in similar outcomes.

In order to capture trends regarding these broad similarities in outcomes, the five SPF ratings were collapsed into three Simplified SPF designations: (a) all schools that received a Red or Orange SPF rating were included in the “Intervention” Simplified SPF designation; (b) all schools that received a Yellow SPF rating were included in the “On Watch” Simplified SPF designation, using the term for Yellow schools employed by the district; and (c) all schools that received a Green or a Blue SPF rating were included in the “High Performing” Simplified SPF
designation. These Simplified SPF categories were then used to run ordinal logit regressions using cubed terms for all the student demographic variables used in Research Question 4, with the exception of the Special Education variable, which only used squared terms as its quadratic coefficient was no longer statistically significant. Like Research Question 5, these regressions also held constant: (a) the percent of teachers that are classified as Fully Qualified, (b) the student-teacher ratio, (c) the number of disciplinary actions that result in instructional loss per 100 students, and (d) dichotomous variables for the academic years 2017-2018 and 2018-2019, with the variable for 2016-2017 being omitted as the reference. Finally, like Research Question 4 these models first explored each individual student demographic predictor. These individual student demographic predictor models were used to produce predicted margins in Stata, which were exported to Excel and then used to create figures in R Studio showing how changes in each student demographic predictor variable related to predicted changes in the probability of receiving each of the three Simplified SPF designation outcomes.

Analysis and Interpretation

Because this is a QuantCrit study, all these results were analyzed and interpreted with explicitly antiracist aims. At no point were disparities in accountability outcomes, school contexts, or student demographics attributed to inherent characteristics of the students themselves or being “caused” by student demographics. Rather, all disparities were framed as reflections of accountability policies and institutional practices, pointing to the need for better policies and practices instead of the need for different kinds of students.

Positionality Statement

One of the reasons I aspired to earn a doctorate was because too often I sat with grieving families and children who had internalized the deficit views that harmful, discriminatory
educational policies and practices beget: mothers who blamed themselves for not speaking English well enough to understand how and why their child was being placed in Special Education; my niece, who tearfully told me she was “too dumb” to pass kindergarten and needed to be retained. Critical research has explored the process and consequences of historically marginalized populations’ internalization of racial hierarchies and the deficit ideologies which maintain them (Kohli, 2014). The interactions I had with families, both as a social worker and a family member, showed me time and again the truly insidious consequences of education policies and practices which teach families and children that they are fundamentally inadequate. This research project was intended for them.

However, as a researcher dedicated to social justice work it is incumbent upon me to name, reflect on, and interrogate how my positionality as a straight, White, cis, English-dominant, middle-class woman represent limitations of this work (Hartsock, 1997; Milner, 2007; North, 2008). This study explores the mechanisms by which accountability policy marginalizes communities that are already marginalized. Yet, because I am both outside these communities as well as privileged rather than disadvantaged by such policies, I may not be able to fully understand them as embodied practices. This work may land in highly personal and painful ways on the historically marginalized communities which this study takes as its focal population, and even new methodologies such as QuantCrit may not be completely adequate for capturing this marginalization. If the findings of such research merely serves to inform, remind, or retraumatize these populations about the various ways that they are socially constructed as inferior or institutionally marginalized, then this study is arguably as destructive as the phenomena it aspires to bring to light. Because of this, in conducting this work I have especially endeavored to not trivialize the experiences of the focal populations of this study, whose identities and experiences
transcend the narrow boundaries that I have employed through demographic categorizations. In doing this work, I lean heavily on my experiences working and living with Mexican, immigrant, and undocumented communities over the last sixteen years, experiences which have engendered a deep place of love and respect in my heart for the families with whom I have been privileged to work and serve. Such feelings are compounded by the love and respect I have for my family who come from similar backgrounds. Despite how my positionality limits my ability to fully understand the issues explored in this dissertation, it is my hope that these limitations are tempered by the attitude of service and love I bring to the project.
Results

Research Question 1: What are the student demographics, EL characteristics, and school contexts per School Performance Framework (SPF) rating bracket?

Schools in the lowest rated brackets consistently served higher proportions of historically marginalized student populations of students of color (SoC), students receiving Free and Reduced Lunch (FRL), Special Education students (SPED), and English Learner students (EL), while serving lower percentages of Gifted and Talented students (GT). Disparities between the lowest and highest SPF ratings (Red and Blue) were the most dramatic. At no point during the study Timeframe did schools in the Blue SPF ratings category serve average student populations that were either (a) above the district average for historically marginalized populations, or (b) below the district average for GT students or Fully Qualified Teachers. The inverse trend was evident regarding schools in the Red SPF ratings category: With the exception of ELs in the 2016-2017 academic year, throughout the study Timeframe Red schools consistently served historically marginalized student populations that exceeded district averages while having percentages of GT students and Fully Qualified Teachers that never reached the district averages. Below, I provide a brief description of the discrepancies between student demographics in the lowest-rated (Red) and highest-rated (Blue) schools as evident in the three-year aggregate means, followed by a summary of general trends between Red and Blue schools for the remaining variable categories. Table 3 shows the means of each variable per SPF ratings bracket for the aggregated three academic years of the study, and Appendix Tables 2, 3, and 4 in Appendix B show the means for each individual academic year. Figure 3 shows the mean percentages of the student demographics and select EL characteristics and school contexts per SPF ratings bracket during each year of the study.
Table 3.
Means of Student Demographics, English Learner Characteristics, Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Years 2016-2017 through 2018-19

<table>
<thead>
<tr>
<th>School Characteristics</th>
<th>Red</th>
<th>Orange</th>
<th>Yellow</th>
<th>Green</th>
<th>Blue</th>
<th>District Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>48</td>
<td>55</td>
<td>186</td>
<td>226</td>
<td>47</td>
<td>562</td>
</tr>
<tr>
<td><strong>%</strong></td>
<td>8.5%</td>
<td>9.8%</td>
<td>33.1%</td>
<td>40.2%</td>
<td>8.4%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Student Demographics**

- Students of Color %: 87.5, 86.2, 78.4, 77.5, 54.7, 77.6
- Free and Reduced Lunch %: 78.5, 76.6, 71.5, 69.0, 42.3, 69.2
- Special Education %: 15.4, 14.4, 12.3, 10.9, 8.3, 11.8
- English Learner %: 36.7, 38.8, 32.2, 37.2, 21.0, 34.3
- Gifted and Talented %: 10.5, 11.6, 12.4, 11.2, 18.7, 12.4

**English Learner Characteristics**

- Special Education as English Learners %: 41.5, 44.0, 35.2, 42.1, 30.3, 39.0
- Spanish-Speaking English Learner %: 87.0, 84.8, 80.3, 78.4, 59.7, 78.8
- English Learners in Gifted and Talented %: 2.1, 2.6, 2.7, 2.7, 10.2, 3.1
- Beginning Level English Learner %: 24.2, 25.6, 22.9, 21.1, 15.7, 22.0
- Intermediate Level English Learner %: 72.2, 68.9, 70.9, 70.0, 66.3, 70.1
- Advanced Level English Learner %: 3.6, 5.4, 6.2, 8.9, 17.9, 7.9

**English Learner Services**

- Redesignation %: 10.5, 9.7, 14.8, 10.4, 18.6, 12.5
- Exit %: 6.7, 6.5, 6.4, 5.7, 10.0, 6.4
- Re-Entry %: 0.7, 1.5, 1.0, 0.7, 1.4, 0.9
- Parent Preference 1 % (bilingual): 40.1, 42.5, 38.6, 41.0, 27.5, 39.1
- Parent Preference 2 % (whatever is at school): 50.5, 49.3, 53.7, 52.9, 64.7, 53.6
- Parent Preference 3 % (nothing): 9.2, 8.5, 7.9, 6.1, 9.4, 7.5
- Mainstream %: 20.6, 35.8, 17.2, 24.2, 34.1, 23.5
- ELA - English %: 69.0, 46.6, 66.9, 55.6, 57.9, 59.9
- ELA - Spanish (ELAS) %: 10.4, 13.9, 15.4, 16.3, 5.7, 14.4
- Dual Language (DL) %: 0.0, 3.7, 0.5, 3.8, 2.2, 2.3
- Native Language (ELAS+DL) %: 10.4, 17.6, 15.9, 20.1, 8.0, 16.7

**School Contexts**

- Total Enrollment: 314.0, 423.3, 479.2, 442.1, 439.2, 441.4
- Student-Teacher Ratio: 14.3, 14.6, 14.6, 15.0, 16.2, 14.9
- Fully Qualified Teacher %: 67.4, 71.4, 80.2, 79.5, 86.1, 78.7
- Disciplinary Actions per 100 Students: 16.5, 11.5, 11.8, 7.6, 6.1, 10.0
- Disciplinary Incidents per 100 Students: 23.2, 18.4, 16.9, 10.5, 8.3, 14.3
- Disciplinary Actions Resulting in Instructional Loss per 100 Students: 11.8, 8.1, 7.0, 4.2, 2.8, 6.1
- Charter School %: 54.2, 43.6, 20.4, 28.8, 31.9, 29.9
Figure 3.
Mean Percentages of Select Student Demographics, EL Characteristics and Services, and School Contexts Across SPF Ratings Brackets for Each Year
Student Demographics

Students of Color: Blue schools served average percentages of students of color that were 22.9 percentage points lower than the district average (m=77.6%) and 32.8 percentage points lower than Red school averages (m=87.5%), meaning that on average students of color populations were 60% larger in Red schools compared to Blue schools. On average, Red schools had 9.9 percentage points more students of color than the district average. Although during the three years of the study all the non-Blue SPF ratings brackets had average populations of students of color that were over 75%, the Blue SPF ratings bracket had average populations of students of color of only 54.7%.

Free and Reduced Lunch (FRL): During the study Timeframe, all non-Blue schools had average FRL populations between 69% and 78.5%, while Blue schools had average FRL populations of only 42.3%. Blue schools served average percentages of FRL students that were 27 percentage points lower than the district average (m=69.2%) and 36.3 percentage points lower than the Red school average (m=78.5%), while Red schools served average FRL populations 9.3 percentage points above the district average. As such, Red schools served average FRL populations that were 85.9% larger than those in Blue schools.

Special Education (SPED): Although it may appear that there was greater parity between Blue schools and the district average regarding SPED populations as Blue schools served an average SPED population that was only 3.5 percentage points below the district average, the general low frequency of these students can be misleading since that 3.5 percentage point discrepancy actually represents an average SPED population that was 38.9% larger than the district average. Similarly, Red schools served average SPED populations that were 3.6 percentage points or 30.1% larger than district average. Taken together, Blue schools served
average SPED populations 7.1 percentage points smaller than those of Red schools. Since the district average for SPED students was 11.8%, a difference of 7.1 percentage points is relatively large, as it translates to Red schools serving average SPED populations (m=15.4%) that were 85.8% larger than those in Blue schools (m=8.3%).

**English Learner (EL):** Red schools did not serve average EL populations that were much larger than the district average as they only surpassed the district average by 2.4 percentage points. However, Blue schools served EL populations that were considerably smaller (m=21.0%) than the district average (m=34.3%), representing a difference of 13.3 percentage points or 38.8% smaller than the district average. Compared to Red schools, Blue schools served EL populations that were 15.7 percentage points smaller, meaning that Red schools served EL populations 74.8% larger than those in Blue schools.

**Gifted and Talented (GT):** The frequency of students receiving GT services is a special point of policy- and practice-based disparities, unless one assumes that talents are not evenly distributed across the population. Rejecting this logic, it is undeniable that the highest- and lowest-rated schools identified gifts and talents in their students at very different rates. On average, nearly one in five students at Blue schools were designated for GT (m=18.7%), while in Red schools only one in ten students (m=10.5%) were designated as such. Both of these rates differ from the district average of 12.4%, with Blue schools’ GT students being about 50% higher than the district average and Red schools’ GT students being 15.2% lower than the district average. Refusing to accept the causal reasoning that there are nearly half as many students with gifts and talents in Red schools as compared to Blue schools, these discrepancies highlight the differential treatment, opportunities, and acknowledgement that students in the majority SoC and FRL Red schools face.
**EL Characteristics and Services**

Unsurprisingly, the discrepancy between the rate of students receiving GT in Blue and Red schools is also evident in the rate of EL participation in GT. About one in ten ELs in Blue schools participated in GT (m=10.2%), while only about one in 50 ELs in Red schools (m=2.1%) were given the same opportunity. Compared to the district average, Blue schools had 228.9% higher rates of ELs in GT. Similarly, Red schools saw a greater percentage of their SPED students co-classified as ELs (m=41.5%), which is 11.2 percentage points higher than the rate in Blue schools (m=30.3%), whose rate of SPED students that are also ELs is 8.7 percentage points or 22.3% less than the district average. A raciolinguistic lens which highlights how Spanish-English bilingualism is especially denigrated in the US might clarify the ideological roots of these discrepancies, as Red schools on average also had 27.4 percentage points or 45.9% larger Spanish-speaking EL populations than Blue schools, whose Spanish-speaking EL populations were 19.2 percentage points smaller than the district average.

Similarly, Red and Blue schools served ELs at markedly different points in their trajectory toward English development. Blue schools on average served percentages of Advanced Level ELs (m=17.9%) that were 10.0 percentage points or 126.1% larger than the district average (m=7.9%). Since the percentage of Intermediate Level ELs were similar between Red and Blue schools (m=72.2% and m=66.3% respectively), the discrepancy of Advanced Level ELs was reflected in a similar discrepancy in the rate of Beginning Level ELs, as Red schools on average had 8.4 percentage points or 53.5% larger Beginning Level EL populations than Blue schools. Since by definition these students are not yet proficient in English, and the Spanish-language version of the primary standardized test used to calculate SPF scores is only available in Spanish until the fourth grade (Colorado Department of Education, n.d.), it is
plausible that some of these students are nonetheless taking standardized tests in English, which might account for the low accountability scores received by the schools with more of these students.

Because Blue schools had higher rates of Advanced Level ELs, it is likewise expected that they also redesignated and exited their ELs from English Learner services at higher rates than the district average (m=12.5% and m=6.4% respectively), whose rates mirrored those of Red schools. Blue schools also Re-Entered ELs into English Learner services afterwards (m=1.4%) at double the rate of Red schools (m=0.7%). These different rates might reflect the kinds of language support services available to students at Red and Blue schools, where on average 13.5 percentage points more students were in Mainstream settings (m=34.1%) than in Red schools (m=20.6%) despite having similar rates of parents choosing the preference option (PPF3) to deny EL services which would make mainstream settings the most appropriate. Interestingly, 12.6 percentage points or 45.8% more parents wanted their EL students to receive native language services (PPF1) in Red schools (m=40.1%) as compared to Blue schools (m=27.5%), although Red schools had no ELs in Dual Language programs.

School Contexts

These differences in EL characteristics and outcomes were sadly also reflected in the percent of teachers the district classifies as “Fully Qualified” to teach ELs, where Red schools had 11.3 percentage points fewer “Fully Qualified” teachers than the district average and Blue schools had 7.4 percentage points more than the district average. As such, when compared to Blue schools Red schools had 18.7 percentage points or 21.7% fewer “Fully Qualified” teachers, despite scholarship that calls for students with the greatest needs to be given more – not less – access to high quality teachers. Not only did Red schools have considerably smaller “Fully
Qualified” teacher populations despite having greater proportions of historically marginalized students, ELs, and Beginning Level ELs, students in these schools also experienced disciplinary regimes unlike those in the Blue schools. During the study Timeframe, on average students in Red schools received 172.8% more disciplinary actions, 181.6% more disciplinary incidents, and 322.5% more disciplinary actions resulting in instructional loss than their peers in Blue schools. Like the GT rates, this study rejects the possibility that students in majority SoC and FRL schools are three times more deserving of discipline that removes them from learning than students in whiter and wealthier schools. These differences in disciplinary environments possibly reflect differences in discipline policies in charter schools, as such schools make up 54.2% of Red schools but only 29.9% of the district.

Summary

During the three years of the study Timeframe, a similar number of schools were categorized as Red (n=48) and Blue (n=47). Despite representing similar proportions of the district, schools at each pole of the SPF ratings brackets served very different kinds of students under very different kinds of school contexts. These findings indicate that the SPF is not only measuring student learning, but also student demographics and, to a lesser extent, school contexts. By measuring conditions extrinsic to student learning, the SPF results in accountability ratings that disadvantage the schools with the most historically marginalized students, Beginning Level ELs, and Spanish-speaking ELs, while not providing these schools with the supports (such as “Fully Qualified” teachers and improved mechanisms for identifying GT students) that these students need and deserve to thrive.
Research Question 2: At what rate did schools remain in, enter into, or exit the most extreme SPF ratings designations of Intervention and Blue status, and what are the student demographics, EL characteristics and services, and school contexts in these statuses?

Although the findings from Research Question 1 indicate that schools with the highest and lowest SPF ratings serve different student populations under different school contexts, a potential counterpoint would be that these differences merely reflect the “education debt” (Ladson-Billings, 2006) owed to historically marginalized students, or the historical failure to provide these students with equitable resources and opportunities that the accountability movement sought to highlight and rectify. As such, the fact that low-rated schools serve higher proportions of historically marginalized students proves the need for accountability policies, which seek to improve outcomes for students by identifying and discouraging low-performance in part by identifying and encouraging high-performance. Research Question 2 sought to investigate this potential counterpoint by exploring the effectiveness of the SPF in achieving its goals of promoting higher performance (and thus higher SPF ratings) while discouraging low performance (and thus low SPF ratings). If it is effective, then schools should demonstrate trends toward higher performance and thus higher ratings over time, while remaining in low-ratings statuses only briefly while the consequences of accountability begin to promote improvements. To address this research question, the two poles of the SPF framework were contrasted: At one end was Intervention Status, representing the schools that earned either the Red or Orange SPF ratings, which indicate school failure and result in district intervention; at the other end, Blue Status represented schools that earned the highest SPF rating possible, and the point toward which the accountability framework should move schools and at which schools should aspire to remain.
Table 4 shows descriptive statistics of the counts and rates of SPF Trends, or schools that remained in, entered into, or exited from these two poles of Intervention Status and Blue Status. An additional row, called “Began in Status,” is included in this table to indicate how many schools were in either Intervention or Blue Status at the beginning of the study; the combined counts of schools that remained in or exited each Status equal the counts of those that “Began in Status.” These results only show counts and rates of the final academic year of the study, 2018-2019, as these data reflect the final outcome of trends at the end of the study Timeframe and using the three-year aggregate would result in repeat counts of schools. In addition, this table disaggregates counts and rates per district-run and charter schools, one of the school context variables identified in the research question. Table 6 shows descriptive statistics of the remaining study variables for schools per each SPF Trend status using the three-year aggregate means in order to capture the average student demographics, EL characteristics and services, and school contexts of schools in each of these SPF Trend statuses throughout the study.

Table 4.
Descriptive Statistics of Schools that Remained In, Entered Into, andExited From Intervention Status and Blue Status per District-Run and Charter Schools as of the Final Year of the Study (2018-2019)

<table>
<thead>
<tr>
<th>SPF Status Trend</th>
<th>District-Run Schools</th>
<th>Charter Schools</th>
<th>All District Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percent (^a)</td>
<td>Count</td>
</tr>
<tr>
<td>Total</td>
<td>129</td>
<td>69.7%</td>
<td>56</td>
</tr>
<tr>
<td><strong>Intervention Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Began in Status</td>
<td>9</td>
<td>56.3%</td>
<td>7</td>
</tr>
<tr>
<td>Remained</td>
<td>4</td>
<td>44.4%</td>
<td>5</td>
</tr>
<tr>
<td>Exit</td>
<td>5</td>
<td>71.4%</td>
<td>2</td>
</tr>
<tr>
<td>Entered</td>
<td>17</td>
<td>54.8%</td>
<td>14</td>
</tr>
<tr>
<td><strong>Blue Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Began in Status</td>
<td>11</td>
<td>61.1%</td>
<td>7</td>
</tr>
<tr>
<td>Remained</td>
<td>4</td>
<td>80.0%</td>
<td>1</td>
</tr>
<tr>
<td>Exit</td>
<td>7</td>
<td>53.9%</td>
<td>6</td>
</tr>
<tr>
<td>Entered</td>
<td>7</td>
<td>87.5%</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\) Percent reflects totals per each Status Trend category
\(^b\) Percent reflects total count of schools in the district (n=185)
SPF Trends: Rates of Schools Remaining In, Entering Into, and Exiting Intervention Status and Blue Status

During the study Timeframe, more than four times more schools entered into Intervention Status (n=31, or 16.8% of schools in the district) than exited it (n=7). Of the 16 schools that began the study in Intervention Status, the majority of them (n=9) remained in this status throughout the three years of the study. Although a similar number of schools began the study in Blue Status (n=18), the majority of them (n=13) consequentially exited it, with only five schools remaining in Blue Status throughout the study and only eight schools entering it. These data suggest that during the study timeframe there was a downward trend in accountability outcomes as schools were more likely to gain and maintain Intervention Status than they were to gain or maintain Blue Status. Similarly, schools were more likely to lose Blue Status than lose Intervention Status. These trends indicate that during the study the SPF accountability framework was ineffective in promoting school success. Together, these trends show that despite the explicit purpose of the SPF accountability framework to promote higher performance and school success – during the study there was not an overall improvement of outcomes in terms of SPF status at the district level but rather schools experienced increasing rates of failure and declining rates of success.

SPF Trends: District-Run and Charter Schools

District-run schools (69.7% of all schools) were overrepresented in the categories of schools that (a) remained in Blue Status (80.0%), (b) entered Blue Status (87.5%), and (c) exited Intervention Status (71.4%), although here their overrepresentation was to a lesser extent. They were underrepresented in the categories of schools that (a) remained in Intervention Status (44.4%), (b) entered Intervention Status (54.8%), and (c) exited Blue Status (53.9%). Charter
schools (30.3% of all schools) were overrepresented in the categories of schools that (a) remained in Intervention Status (43.8%), and (b) entered Intervention Status (55.6%). They were underrepresented in the categories of schools that (a) exited Intervention Status (28.6%), (b) remained in Blue Status (20.0%), and (c) entered Blue Status (12.5%). These data suggest that during the study Timeframe charter schools were less likely than district-run schools to achieve the high performance that the accountability framework seeks to promote, as only 1.8% of charters entered into or remained in Blue Status, while 3.1% of district-run schools remained in Blue Status and 5.4% of them entered into it. At the same time, charters were twice as likely to exit Blue Status than district-run schools, with 10.7% of charters and only 5.4% of district-run schools exiting. Charters remained in Intervention Status at a rate (8.9%) almost three times higher than district-run schools (3.1%). Similarly, during the study Timeframe one in every four charters (25.0%) entered into Intervention Status, while only 13.2% of district-run schools did the same. Together, these data show that during the study Timeframe charters were more likely to be or become low performing than district-run schools while being less likely to be or become high performing.

SPF Trends: Student Demographic, EL Characteristics and Services, and School Contexts

These discrepancies might be partially attributable to the different student populations and school contexts in the schools in each of the SPF Trend statuses (Table 6). Table 5 (below) presents a key to the abbreviated table variable labels used in Table 6 and elsewhere in this chapter.
### Table 5. Key To Abbreviated Variable Names

<table>
<thead>
<tr>
<th>Abbreviated Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student Demographics</strong></td>
<td></td>
</tr>
<tr>
<td>SoC %</td>
<td>Percent of students that are students of color (SoC)</td>
</tr>
<tr>
<td>FRL %</td>
<td>Percent of students that receive Free and Reduced Lunch (FRL)</td>
</tr>
<tr>
<td>SPED %</td>
<td>Percent of students that receive Special Education (SPED) services</td>
</tr>
<tr>
<td>EL %</td>
<td>Percent of students that receive English Learner (EL) services</td>
</tr>
<tr>
<td>GT %</td>
<td>Percent of students that receive Gifted and Talented services (GT)</td>
</tr>
<tr>
<td><strong>English Learner Characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>SPED as ELs %</td>
<td>Percent of Special Education students that are also ELs</td>
</tr>
<tr>
<td>Spanish EL %</td>
<td>Percent of ELs students that are Spanish speakers</td>
</tr>
<tr>
<td>ELs in GT %</td>
<td>Percent of ELs students that receive Gifted and Talented services</td>
</tr>
<tr>
<td>Beginning EL %</td>
<td>Percent of ELs that are in the Beginning Level of English acquisition</td>
</tr>
<tr>
<td>Intermediate EL %</td>
<td>Percent of ELs that are in the Intermediate Level of English acquisition</td>
</tr>
<tr>
<td>Advanced EL %</td>
<td>Percent of ELs that are in the Advanced Level of English acquisition</td>
</tr>
<tr>
<td><strong>English Learner Services</strong></td>
<td></td>
</tr>
<tr>
<td>Redes. %</td>
<td>Rate at which English Learners were redesignated from EL services</td>
</tr>
<tr>
<td>Exit %</td>
<td>Rate at which English Learners were exited from EL services</td>
</tr>
<tr>
<td>Re-Entry %</td>
<td>Rate at which English Learners were re-entered from EL services</td>
</tr>
<tr>
<td>PP1 %</td>
<td>Percent of families who request native language supports designed for ELs for their EL children</td>
</tr>
<tr>
<td>PP2 %</td>
<td>Percent of families who request English-only supports designed for ELs for their EL children</td>
</tr>
<tr>
<td>PP3 %</td>
<td>Percent of families who request no supports for their EL children</td>
</tr>
<tr>
<td>Main. %</td>
<td>Percent of ELs placed in Mainstream programs, which are not specifically designed for ELs</td>
</tr>
<tr>
<td>ELA-E %</td>
<td>Percent of ELs in English Language Acquisition-English programs, which are specifically designed for ELs and taught through English-only</td>
</tr>
<tr>
<td>ELA-S %</td>
<td>Percent of ELs placed in English Language Acquisition-Spanish programs, which are specifically designed for ELs and taught through Spanish</td>
</tr>
<tr>
<td>DL %</td>
<td>Percent of ELs placed in Dual Language programs</td>
</tr>
<tr>
<td>Nat. Lang. %</td>
<td>Percent of ELs placed in either Dual Language or ELA-S programs</td>
</tr>
<tr>
<td><strong>School Contexts</strong></td>
<td></td>
</tr>
<tr>
<td>Enrollment</td>
<td>Total student enrollment</td>
</tr>
<tr>
<td>Student-Teacher Ratio</td>
<td>Ratio of students to teachers</td>
</tr>
<tr>
<td>Full. Qual. Teacher %</td>
<td>Percent of teachers with the label of “Fully Qualified” to teach emergent bilingual students according to district metrics</td>
</tr>
<tr>
<td>Disp. Actions Rate</td>
<td>Count of disciplinary actions per 100 students</td>
</tr>
<tr>
<td>Disp. Incidents Rate</td>
<td>Count of disciplinary incidents per 100 students</td>
</tr>
<tr>
<td>Disp. Instruction Loss Rate</td>
<td>Count of disciplinary actions that result in instructional loss per 100 students</td>
</tr>
<tr>
<td>SPF %</td>
<td>Percent of SPF points earned out of total points possible</td>
</tr>
</tbody>
</table>
Table 6.
Descriptive Statistics of Means of Schools that Remained In, Entered Into, and Exited From Intervention Status and Blue Status Across the Three-Year Study Timeframe Aggregate

<table>
<thead>
<tr>
<th></th>
<th>Intervention Status</th>
<th>Blue Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Remain</td>
<td>Enter</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>%</td>
<td>4.9%</td>
<td>16.8%</td>
</tr>
</tbody>
</table>

Student Demographics
- SoC %: 87.5, 87.3, 80.0, 29.5, 59.0, 73.9
- FRL %: 75.8, 77.6, 74.9, 15.5, 48.5, 62.9
- SPED %: 18.2, 12.7, 12.0, 5.7, 10.2, 10.5
- EL %: 31.9, 40.6, 44.7, 9.8, 21.9, 34.0
- GT %: 10.1, 12.0, 8.5, 23.1, 12.9, 16.3

English Learner Characteristics
- SPED as ELs %: 39.1, 46.4, 40.9, 11.8, 38.1, 44.4
- Spanish EL %: 83.7, 88.3, 80.9, 36.7, 62.4, 77.3
- ELs in GT %: 3.9, 1.9, 2.4, 28.9, 6.0, 2.2
- Beginning EL %: 27.8, 23.6, 24.6, 6.4, 15.9, 20.8
- Intermediate EL %: 67.8, 70.5, 69.7, 66.5, 70.4, 69.5
- Advanced EL %: 4.4, 5.9, 5.7, 27.1, 13.7, 9.7

English Learner Services
- Redes. %: 9.3, 9.8, 10.3, 27.2, 9.6, 25.9
- Exit %: 4.2, 6.0, 4.8, 14.4, 7.1, 7.4
- Re-Entry %: 0.6, 1.3, 1.4, 2.8, 0.4, 1.0
- PP1 %: 37.5, 44.1, 49.5, 11.5, 34.8, 37.7
- PP2 %: 51.5, 47.9, 45.6, 75.8, 63.7, 54.8
- PP3 %: 12.0, 7.8, 4.9, 12.6, 3.9, 8.6
- Main. %: 38.3, 34.6, 21.2, 34.0, 13.9, 40.6
- ELA-E %: 46.7, 49.2, 49.5, 66.0, 86.1, 43.7
- ELA-S %: 11.1, 12.9, 29.3, 0.00, 0.00, 9.3
- DL %: 3.8, 3.3, 0.00, 0.00, 0.00, 6.4
- Nat. Lang. %: 15.0, 16.2, 29.3, 0.00, 0.00, 15.7

School Contexts
- Enrollment: 228.4, 419.4, 545.7, 459.1, 473.5, 369.0
- Students-Teacher Ratio: 13.7, 14.6, 14.7, 18.0, 15.6, 14.9
- Full. Qual. Teacher %: 61.3, 75.9, 78.4, 88.8, 81.8, 82.9
- Disp. Actions Rate: 17.7, 18.8, 20.6, 5.5, 15.1, 7.4
- Disp. Incidents Rate: 11.7, 13.3, 14.7, 4.5, 10.7, 5.5
- Disp. Instructional Loss Rate: 10.0, 8.4, 9.3, 2.5, 2.8, 3.2

Just as the findings from Research Question 1 showed that the highest- and lowest-rated schools vary across student demographics, EL characteristics and services, and school contexts, so, too, do the trends of schools that remained in, entered into, and exited from the extremes of the SPF ratings brackets — Intervention Status and Blue Status — vary along these lines, with the
most distinct variation evidenced between the student populations in the schools that were consistently low- and high-rated. Compared to schools that remained in Blue Status for every year of the study, schools that remained in Intervention Status on average had nearly three times larger proportions of students of color (87.5%), EL students (31.9%), and SPED students (18.2%), and five times larger FRL populations (75.8%), with less than half the rate of GT students (10.1%). Schools that were able to exit Intervention Status had smaller average proportions of SoC (80.0%) and SPED students (12.0%) than schools that remained in this status. Interestingly, schools that exited Intervention Status had larger proportions of ELs (12.8 percentage points more) than schools that remained. Schools that entered Intervention Status not only had larger proportions of EL students than schools that remained in Intervention Status, but they also had slightly larger FRL proportions and almost identical SoC proportions. These data imply that schools that entered Intervention Status had historically marginalized student populations that were similar to those that remained in Intervention status, while schools that exited this status had smaller proportions of these students with the exception of ELs.

Likewise, the 13 schools that exited Blue Status had much larger average proportions of SoC (73.9%), FRL (62.9%), SPED (10.5%) and EL (34.0%) students than schools that remained in Blue Status, whose average proportions of SoC (29.5%), FRL (15.5%), SPED (5.7%), and EL (9.8%) students were approximately one-quarter to one half as large. Schools that entered Blue Status had average historically marginalized student populations that were somewhat in the middle of these two, with SoC, FRL, and EL populations that were respectively 29.5 (SoC), 33 (FRL), and 12.8 (EL) percentage points higher than schools that remained in Blue Status, but 13.9 (SoC), 14.4 (FRL), and 12.1 (EL) percentage points lower than schools that exited it. Just as with the trends evident in the Intervention Status, these data indicate that schools with larger
proportions of historically marginalized students were more likely to exit Blue Status, while schools with smaller proportions of these students were more likely to enter into or remain in it.

Similar disparities were evident in EL characteristics and services, as schools that experienced Intervention status had a little more than twice the proportion of Spanish-speaking ELs and approximately four times higher rates of ELs in SPED than schools that remained in Blue Status. Notably, all schools in this analysis had low rates of ELs in GT, ranging between 1.9% and 6%, except for schools that remained in Blue Status, where nearly one in three GT students were also ELs. However, these schools also served five to six times larger proportions of Advanced Level ELs as compared to all schools that experienced Intervention Status – whose Beginning Level EL populations likewise were about four times larger. The larger proportion of Advanced Level ELs in schools that remained in Blue Status is mirrored in these schools’ higher rates of redesignating and exiting students from English Learner services. These rates appear to be divorced from the kinds of program settings EL students were in as these schools had very similar rates of Mainstream participation as those that experienced Intervention Status, with differences in ELA-E and ELA-S appearing to correspond to differences in Spanish-language ELs and parent preferences.

While ELs seemed to be placed in program settings following similar logics in both Intervention and Blue Status schools, in schools that remained in Blue Status 88.8% of their teachers were Fully Qualified to work with such students while only 61.3% of teachers were similarly qualified in the schools that remained in Intervention Status. The difference in teacher quality was compounded by differences in disciplinary environments, as students in schools that experienced Intervention Status received disciplinary actions and incidents between twice to five times more frequently than students in schools that remained in Blue Status. Sadly, students in
schools that remained in Intervention Status received four times higher rates of disciplinary actions that resulted in instructional loss than their peers in schools that remained Blue.

Summary

Together, these trends indicate that schools with greater proportions of students of color, students receiving Free and Reduced Lunch, Special Education students, Spanish-speaking ELs, and Beginning Level ELs in addition to higher rates of discipline were overrepresented in schools that remained in Intervention Status, while having fewer proportions and rates of these metrics was evident in schools that exited Intervention Status. Conversely, schools that remained in Blue Status had strikingly smaller proportions of these students and lower rates of discipline, while having larger proportions of these students and higher rates of discipline was evident in schools that exited Blue Status. Schools that experienced Blue Status at some point during the study Timeframe all had higher rates of Fully Qualified teachers and Gifted and Talented students with lower rates of discipline than schools that experienced Intervention Status. Further during the study timeframe there was a downward trajectory of more schools entering into or remaining in Intervention Status than entering into or remaining in Blue Status. Together these findings indicate that the SPF was not successful in promoting higher degrees of school success. This leaves future research questions about the impact of not incorporating the student demographic, EL characteristic and services, and school context discrepancies explored into the accountability framework.

Research Question 3: What are the student demographics, EL characteristics and services, and school contexts per charters and district-run schools?

The findings from the previous two research questions have indicated that student demographics, EL characteristics and services, and school contexts vary both across SPF ratings
brackets as well as across schools that remained in, entered into, and exited from the SPF statuses that represent the special focus of the accountability framework. Low accountability ratings can lead to school closure and replacement with restart by a charter school. Additionally, these previous findings have indicated that more historically marginalized students are learning in distinct school contexts in these low-rated schools. Thus this third research question sought to understand whether these same metrics also vary between district-run schools and the charters that are potentially replacing them. Table 7 shows the means of the study variables in each academic year as well as the three-year aggregate means. The same abbreviated variable names used in the previous section are employed in Table 7; refer to Table 5 for a description of variable names.

**Table 7.**
*Means of Study Variables per District-Run And Charter Schools for Each Year of Study and Three-Year Aggregate*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>132</td>
<td>54</td>
<td>133</td>
<td>58</td>
</tr>
<tr>
<td><strong>%</strong></td>
<td>71.0%</td>
<td>29.0%</td>
<td>69.6%</td>
<td>30.4%</td>
</tr>
</tbody>
</table>

**Student Demographics**
- SoC % 74.9 84.4 74.7 85.1 74.4 85.0 74.7 84.9
- FRL % 67.1 74.1 67.6 74.9 66.1 74.5 66.9 74.5
- SPED % 11.1 11.0 12.1 12.1 12.4 12.3 11.9 11.8
- EL % 32.3 37.3 33.5 40.3 31.7 39.2 32.5 39.0
- GT % 12.4 15.1 15.0 14.4 9.3 10.0 11.8 13.1

**English Learner Characteristics**
- SPED as ELs % 39.5 45.8 34.7 45.2 34.8 45.9 36.3 45.6
- Spanish EL % 76.6 84.3 76.0 86.1 75.6 85.8 76.1 85.4
- ELs in GT % 3.3 3.6 2.8 2.3 3.5 2.2 3.2 2.8
- Beginning EL % 22.2 10.3 24.7 17.2 26.7 21.0 24.5 16.3
- Intermediate EL % 68.1 80.4 67.9 77.9 65.0 73.1 67.0 77.0
- Advanced EL % 9.7 9.3 7.4 4.9 8.2 5.9 8.4 6.6

**English Learner Services**
- Redes. % 7.8 7.8 12.3 15.1 14.4 21.4 11.5 15.0
- Exit % 3.6 15.5 4.6 7.2 6.3 9.1 4.8 10.4
Re-Entry %  0.9  1.2  0.3  0.4  1.3  1.8  0.8  0.9  1.1
PP1 %  39.5  37.5  39.3  37.5  39.8  39.8  39.5  38.3
PP2 %  54.3  50.2  54.2  52.9  54.4  53.2  54.3  52.1
PP3 %  6.5  11.5  6.4  8.1  8.0  6.5  6.7  8.8
Main. %  3.4  95.5  4.5  82.0  3.2  32.7  3.7  69.4
ELA-E %  74.8  3.8  71.9  16.8  74.4  60.6  73.7  27.7
ELA-S %  18.7  0.7  20.6  1.2  18.9  6.7  19.4  2.9
DL %  3.1  0.0  3.1  0.0  3.5  0.0  3.2  0.0
Nat. Lang. %  21.8  0.7  23.6  1.2  22.4  6.7  22.6  2.9

School Contexts

Enrollment  495.1  330.9  479.8  340.2  481.4  345.3  485.4  339.0
Student-Teacher Ratio  15.5  14.4  15.2  13.6  15.2  13.6  15.3  13.8
Full. Qual. Teacher %  82.8  25.0  84.4  44.2  81.6  No data  83.0  42.6
Disp. Actions Rate  9.6  22.1  13.5  19.2  12.9  18.6  12.0  19.9
Disp. Incidents Rate  6.9  15.9  9.0  14.0  9.1  11.8  8.3  13.9
Disp. Instruction Loss Rate  6.4  12.2  4.2  7.5  3.9  7.8  4.8  9.1
SPF %  57.6  58.2  55.2  50.1  51.9  48.5  54.9  52.2

In each year of the study, when compared to district-run schools charters served higher percentages of students of color, FRL students, ELs, Spanish-speaking ELs, Special Education students that are ELs, and Intermediate Level ELs, with lower percentages of Beginning and Advanced Level ELs. In some cases, these differences were stark. For example, when compared to district-run schools on average charters served 13.7% (or 10.2 percentage points) larger proportions of students of color, 20.0% (6.5 percentage points) larger proportions of ELs, and 25.6% (9.3 percentage points) larger proportions of SPED students that were ELs, with 33.5% (8.2 percentage points) smaller proportions of Beginning Level ELs and 21.4% (1.8 percentage points) smaller proportions of Advanced Level ELs. However, there were no consistent disparities between charters and district-run schools regarding proportions of GT students or ELs in GT, and they each served nearly identical proportions of SPED students. These data indicate that on average charters served student populations that were less White, less wealthy, and more bilingual than district-run schools, with higher rates of Special Education students that were ELs,
and ELs that were Spanish-speakers and Intermediate Level than district-run schools. These findings correspond with the previous results that found charters to be overrepresented in the schools experiencing Intervention Status (Research Question 2), and historically marginalized students overrepresented in low-rated schools (Research Question 1).

However, these larger proportions of ELs that would benefit from and qualify for the native language programming offered in the district did not translate to charters offering more students such opportunities. Despite district-run schools and charters having similar percentages of families preferring native language supports (PPF1) and English-only language instruction designed for ELs (PPF2), charters had considerably higher rates of placing ELs in Mainstream class settings, which reflect neither of these preferences. While on average only 3.7% of ELs in district-run schools were placed in Mainstream programming, 69.4% of ELs in charter schools were in Mainstream classes, representing a 1,775.7% higher rate of Mainstream placement. Charters placed approximately one in four (27.7%) ELs in English-only language instruction designed for ELs, or ELA-English classes, while district-run schools placed ELs in these programs at nearly three times that rate (73.7%). Unfortunately, this trend is also evident regarding native language programming, as on average nearly one in five (19.4%) ELs in district-run schools were placed in ELA-Spanish programs with only 2.9% of ELs in charters given the same opportunity. Despite these dramatic differences in participation rates in programs designed for emergent bilingual students, charter schools exited ELs from English Language services 116.7% more frequently than district-run schools, exiting on average one in ten ELs every year, while district-run school exited about one out of every 25 ELs annually. Although charters also had 37.5% higher average re-entry rates for these students, indicating that exiting them was premature, this higher rate of re-entries does not account for the discrepancies of
removing ELs from English Language supports, leading to the question of what happens to these students in charters once they are no longer tracked as ELs.

Students in charters were not only less likely to participate in programs designed for emergent bilinguals, they were also much more likely to experience disciplinary actions and incidents and lose instructional time as a consequence. In every year of the study, charters had higher rates of disciplinary actions, disciplinary incidents, and disciplinary actions resulting in instructional loss than district-run schools, with 65.8% higher rates of disciplinary actions, 67.5% higher rates of disciplinary incidents, and nearly double (89.6%) higher rates of disciplinary actions resulting in instructional loss. Charters also had considerably smaller percentages of their teachers that are fully qualified to work with emergent bilingual students according to the district metrics of “Fully Qualified” teacher, with rates of Fully Qualified teachers that were approximately one-third to one-half of those in district-run schools. However, these data were incomplete as the district did not report on charters’ Fully Qualified teacher rates in the 2018-2019 academic year and only partial reporting was available for the previous years.

**Summary**

Taken together, this analysis indicates that despite having greater proportions of students of color, students receiving Free and Reduced Lunch services, English Learners, and the Intermediate Level and Spanish-speaking English Learners that would especially benefit from native language supports or other programs specifically designed to serve ELs, charters provided dramatically fewer of these opportunities to their ELs. Further all students in charters were subject to stricter disciplinary environments. While it might be argued that high rates of Mainstream classes for ELs, exiting ELs from English Learner status, and discipline are all reflections of the different approaches to education that make charters unique and successful
alternatives to district-run schools, that success was not evidenced in SPF scores, as charters’ average SPF scores were a 4.9% lower than those in the district.

**Research Question 4: Do student demographics predict percent SPF points earned?**

The previous findings have consistently indicated that student demographics, EL characteristics and services, and school contexts all vary between schools with high- and low-SPF ratings, whether one evaluates this variation across SPF ratings brackets, schools that remained in, began in, or ceased being especially high- or low-performing over time, or through the charter or district school statue. However, none of these analyses have included tests of statistical significance, meaning that despite the consistency of these variations the differences could just be statistical “noise,” or random fluctuations above and below the means.

To evaluate whether these differences are statistically significant or just random, I ran a series of OLS regressions predicting the percent of SPF points earned per student demographic while controlling for (a) the percentage of teachers classified as “Fully Qualified” (F.Tch %), (b) the student-teacher ratio (ST Ratio), and (c) the rate of disciplinary actions resulting in instructional loss per 100 students (Disp. Loss). Because these models represent the three-year aggregated data, they also controlled for year by using dichotomous variables for the academic years of 2017-2018 (Year 2017) and 2018-2019 (Year 2018), with the 2016-2017 academic year variable omitted as the reference. Table 8 shows the results from the series of individual student demographic predictors, along with saturated models in which all student demographic predictors are included (and thus controlled for). For each student demographic predictor, a regression model is shown using linear and then polynomial terms, which allow for curvilinear relationships between the predictor and the outcome. Quadratic terms are indicated with a
squared exponent next to the variable name, and cubed terms are indicated with a cubed exponent next to the variable name.

*Individual Student Demographic Predictors*

**Predictor: Student of Color (SoC) Percent**

In Models 1 and 2, the predictor is the percentage of a school population that are classified as students of color (SoC %). Model 1 uses a linear term for the SoC predictor, while Model 2 uses cubed terms. The quadratic term (SoC %²) is statistically significant, indicating that the additional cubed term (SoC %³) was appropriate, with the slightly higher Adjusted R² (Adj R²) value in Model 2 compared to Model 1 likewise indicating that Model 2 was a better fit for these data, as Model 1 accounted for 29% of the variation of the data while Model 2 accounted for 33% of the variation. In both Model 1 and Model 2, the percentage of students of color is a statistically significant predictor of SPF score even when controlling for the percentage of Fully Qualified teachers, student-teacher ratio, rate of disciplinary actions resulting in instructional loss, and year. The coefficient on the percent students of color variable in the cubed model (Model 2) indicates that, for every one point positive difference in the percent of students that are students of color in a school (in layman’s terms, schools serving one percentage point more of SoC are predicted to have scores that are 2.13 SPF points lower, on average), even when holding constant all of the control variables we would expect to see 2.13 fewer SPF percentage points. Model 2 was statistically significant (R² = [0.33], F(8, 421) = [27.01], p = [0.00]). The percent of students of color in a school significantly predicted the percent of SPF points earned (β = [-2.13], p = [0.002]), and its the fitted model was:

\[
\text{Percent SPF Points Earned} = 107.72 + 2.13(\text{percent students of color}^3) + 0.02(\text{percent Fully Qualified teachers}) + 0.4(\text{student-teacher ratio}) + 0.29(\text{rate of disciplinary actions resulting in instructional loss}) + 2.90(\text{Year 2017}) + 6.27(\text{Year 2018})
\]
Table 8.
Individual Predictor and Saturated Models OLS Regressions with Cubed Terms for Academic Years 2016-2017 Through 2018-2019

<table>
<thead>
<tr>
<th>Predictor</th>
<th>SoC</th>
<th>EL</th>
<th>SPED</th>
<th>FRL</th>
<th>Saturated: SoC</th>
<th>Saturated: FRL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>SoC %</td>
<td>-0.25***</td>
<td>-2.13**</td>
<td></td>
<td></td>
<td>-0.33***</td>
<td>-2.00**</td>
</tr>
<tr>
<td>SoC % ^2</td>
<td>0.03*</td>
<td></td>
<td></td>
<td></td>
<td>0.03*</td>
<td></td>
</tr>
<tr>
<td>SoC % ^3</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td>0.00+</td>
<td></td>
</tr>
<tr>
<td>EL %</td>
<td></td>
<td>-0.11***</td>
<td>-1.59***</td>
<td></td>
<td>0.14**</td>
<td>0.07</td>
</tr>
<tr>
<td>EL % ^2</td>
<td></td>
<td>0.04***</td>
<td></td>
<td></td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>EL % ^3</td>
<td></td>
<td>0.00***</td>
<td></td>
<td></td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>SPED %</td>
<td></td>
<td>-0.76***</td>
<td>-8.39***</td>
<td></td>
<td>-0.47**</td>
<td>-3.27+</td>
</tr>
<tr>
<td>SPED % ^2</td>
<td></td>
<td>0.50***</td>
<td></td>
<td></td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>SPED % ^3</td>
<td></td>
<td>-0.01***</td>
<td></td>
<td></td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>FRL %</td>
<td></td>
<td></td>
<td></td>
<td>-0.21***</td>
<td>-1.55***</td>
<td>-0.29***</td>
</tr>
<tr>
<td>FRL % ^2</td>
<td></td>
<td></td>
<td></td>
<td>0.02***</td>
<td></td>
<td>0.02**</td>
</tr>
<tr>
<td>FRL % ^3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00*</td>
<td>0.00*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Controls</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F.Tch %</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>ST Ratio</td>
<td>0.39</td>
<td>0.44+</td>
<td>1.11***</td>
<td>0.81**</td>
<td>0.79**</td>
</tr>
<tr>
<td></td>
<td>Disp. Loss</td>
<td>-0.32***</td>
<td>-0.29***</td>
<td>-0.36***</td>
<td>-0.29***</td>
<td>-0.33***</td>
</tr>
<tr>
<td></td>
<td>Year 2017</td>
<td>-3.13*</td>
<td>-2.90*</td>
<td>-3.31*</td>
<td>-2.94+</td>
<td>-2.65+</td>
</tr>
<tr>
<td></td>
<td>Year 2018</td>
<td>-6.55***</td>
<td>-6.27***</td>
<td>-7.11***</td>
<td>-7.02***</td>
<td>-5.41**</td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>70.08***</td>
<td>107.72***</td>
<td>44.31***</td>
<td>61.98***</td>
<td>52.78***</td>
</tr>
<tr>
<td></td>
<td>Adj R2</td>
<td>0.29</td>
<td>0.33</td>
<td>0.20</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Obv.</td>
<td>430</td>
<td>430</td>
<td>416</td>
<td>416</td>
<td>421</td>
</tr>
</tbody>
</table>

+ Indicates p-value ≤ 0.1
* Indicates p-value ≤ 0.05
** Indicates p-value ≤ 0.01
*** Indicates p-value ≤ 0.001

Note: Year variables represent binaries for the academic years 2017-2018 and 2018-2019; the binary variable for the academic year 2016-2017 was omitted as the reference.
Predictor: English Learner (EL) Percent

Models 3 and 4 use the percentage of English Learners in a school as the predictor, with the statistically significant quadratic and cubed terms indicating a curvilinear relationship between the percent of SPF points earned and the percent of students classified as English Learners, and a slightly higher Adjusted R2 indicating that the cubed model is the better fit as it accounted for 26% of the variation in the data while Model 3 only accounted of 20%. The percent of students that are English Learners is a statistically significant predictor of the percent of SPF points earned in which SPF points earned are predicted to be 1.59 lower, on average, with every one point positive difference in the percent of students in a school that are ELs when holding constant the controls. In other words, Model 4 was statistically significant ($R^2 = 0.26$, $F(8, 407) = 19.26$, $p = 0.00$), with the percent of English Learners in a school significantly predicting the percent of SPF points earned ($\beta = -1.59$, $p = 0.000$). Its fitted model was:

$$\text{Percent SPF Points Earned} = 61.98 - 1.59(\text{percent English Learners}^3) + 0.02(\text{percent Fully Qualified teachers}) + 0.81(\text{student-teacher ratio}) - 0.29(\text{rate of disciplinary actions resulting in instructional loss}) - 2.94(\text{Year 2017}) - 7.02(\text{Year 2018})$$

Predictor: Special Education (SPED) Percent

Models 5 and 6 use the percentage of students receiving Special Education services as the predictor. Like in Model 4, the statistically significant quadratic and cubed terms indicate a curvilinear relationship between the percent of SPF points earned and the percent of students classified as Special Education, and a slightly higher Adjusted R2 indicating that the cubed model is the better fit as it accounts for 24% of the variation in the data while Model 5 only accounts of 21%. Like the previous models, in Model 6 the percent of students that receive Special Education services is a statistically significant predictor of percent SPF points earned
in which SPF points earned are predicted to be 8.39 lower, on average, with every one point positive difference in the percent of students in a school that are SPED when holding constant the controls. Model 6 was statistically significant (R2 = [0.24], F(8, 412) = [17.69], p = [0.00]), with the percent of students in Special Education in a school significantly predicting the percent of SPF points earned (β = [-8.39], p = [0.000]). Its fitted model was:

Percent SPF Points Earned = 87.85 + -8.39(percent Special Education students³) + 0.04(percent Fully Qualified teachers) + 0.69(student-teacher ratio) + -0.32(rate of disciplinary actions resulting in instructional loss) + -2.13(Year 2017) + -5.28(Year 2018)

Predictor: Free and Reduced Lunch (FRL) Percent

Models 7 and 8 use the percentage of students receiving Free and Reduced Lunch services as the predictor. Like in Models 4 and 6, the statistically significant quadratic and cubed terms indicate a curvilinear relationship between the percent of SPF points earned and the percent of students classified as Free and Reduced Lunch, and a slightly higher Adjusted R2 indicating that the cubed model is the better fit as it accounts for 34% of the variation in the data while Model 7 only accounts of 28%. Like the previous models, in Model 8 the percent of students that receive Free and Reduced Lunch services is a statistically significant predictor of the percent SPF points earned in which SPF points earned are predicted to be 1.55 lower, on average, with every one point positive difference in the percent of students in a school that are FRL when holding constant the controls. Model 8 was statistically significant (R2 = [0.34], F(6, 423) = [28.59], p = [0.00]), with the percent students receiving Free and Reduced Lunch in a school significantly predicting the percent SPF points earned (β = [-1.55], p = [0.000]). Its fitted model was:

Percent SPF Points Earned = 85.53 + -1.55(percent Free and Reduced Lunch students³) + 0.02(percent Fully Qualified teachers) + 0.26(student-teacher ratio) + -0.32(rate of disciplinary actions resulting in instructional loss) + -3.19(Year 2017) + -6.67(Year 2018)
Saturated Models

Two sets of saturated models were also created to test the predictive power of these student demographic variables when they were combined into single models. Because of the high degree of collinearity ($r=0.95$) between the percent student of color and the percent Free and Reduced Lunch variables discussed in the Methods section, they could not be used together in a single model. For this reason, one saturated model includes the percent students of color variable, and the other includes the percent Free and Reduced Lunch variable. Like the previous models, these saturated models are presented in the table first using linear and then cubed terms.

Saturated Model with Student of Color Percent Variable

Only the percent students of color variable continued to be a statistically significant predictor in the saturated model using cubed terms (Model 10), although all the student demographic variables were statistically significant predictors of the percent of SPF points earned in the previous individual models and the saturated model using linear terms (Model 9). This indicates that in the previous models, the variables that appeared to be statistically significant predictors of the percent SPF points earned (e.g., percent English Learners and percent Special Education) perhaps were not predictors in and of themselves but rather their significance was derived from co-occurring characteristics of these students; namely, that students of color are overrepresented in the English Learner and Special Education classifications. In this way, the EL and SPED variables appeared to be significant predictors of the percent SPF points earned, but this relationship was not due to students’ EL and SPED status but rather students’ co-occurring student of color status. Because of this, when all these student demographic predictors were included in a single saturated model, we see that only the percent students of color variable continues to be significant, as this was the classification that was most
responsible for the relationship with the percent SPF points earned. Like in previous model sets, the model using curvilinear terms (Model 10) has a slightly higher R2 than the model using linear terms (Model 9), indicating that the curvilinear model – in which only the percent student of color variable was a statistically significant predictor of the percent SPF points earned – is a better fit for these data. Model 10 was statistically significant (R2 = [0.35], F(14, 394)) = [16.68], p = [0.00]), with the percent students of color in a school significantly predicting the percent SPF points earned (β = [-2.00], p = [0.006]). Its fitted model (Model 10) was:

Percent SPF Points Earned = 122.33 + -2.00(percent students of color³) + 0.07(percent English Learners³) + -3.27(percent Special Education students³) + 0.02(percent Fully Qualified teachers) + 0.17(student-teacher ratio) + -0.25(rate of disciplinary actions resulting in instructional loss) + -2.89(Year 2017) + -6.34(Year 2018)

Saturated Model with Free and Reduced Lunch Percent Variable

Similarly, although all the student demographic variables were statistically significant predictors of the percent of SPF points earned in the previous individual models and the saturated model using linear terms (Model 11), when combined into a single saturated model using cubed terms (Model 12), only the variable for the percent of Free and Reduced Lunch students continued to be a statistically significant predictor. Like the other saturated model, this indicates that the predictive power of the variables which previously appeared to be statistically significant predictors of the percent SPF points earned (e.g., percent English Learners and percent Special Education) was perhaps derived from co-occurring characteristics in which many of these students were also classified as receiving Free and Reduced Lunch. As such, when all these student demographic predictors were included in a single saturated model, we see that only the percent FRL variable continues to be significant, as this was the classification that was most responsible for the relationship with the percent SPF points earned. Continuing the trend evident throughout these regressions, the model using curvilinear terms (Model 12) has a slightly higher
R2 than the model using linear terms (Model 11), indicating that the curvilinear model – in which only the percent FRL variable was a statistically significant predictor of the percent SPF points earned – is a better fit for these data. Model 12 was statistically significant ($R^2 = [0.35]$, $F(14, 394) = [16.95], p = [0.00])$, with the percent Free and Reduced Lunch students in a school significantly predicting the percent SPF points earned ($\beta = [-1.38], p = [0.000]$). Its fitted model (Model 12) was:

\[
\text{Percent SPF Points Earned} = 122.33 + -1.38(\text{percent Free and Reduced Lunch students}^3) + 0.21(\text{percent English Learners}^3) + -1.17(\text{percent Special Education students}^3) + 0.03(\text{percent Fully Qualified teachers}) + 0.08(\text{student-teacher ratio}) + -0.24(\text{rate of disciplinary actions resulting in instructional loss}) + -3.02(\text{Year 2017}) + -6.65(\text{Year 2018})
\]

**Figures of Predicted Percent SPF Points Earned**

Figure 4 shows panels of the predicted percent of SPF points earned per student demographic variable. These figures were created using the individual predictor models using curvilinear terms discussed in detail above (Model 2, Model 4, Model 6, and Model 8). The y-axis shows the predicted percent of SPF points earned, and the x-axis shows the corresponding change in each student demographic variable. The range of x-axis values begin with 0% and continues through the 99th percentile of each student demographic in order to capture the proportions of student demographic populations as they existed in the district with the exception of the students of color variable, which begins with 20% as its first percentile value was 18.1 and its 50th percentile value was 90.0 (Table 9).
Table 9.
Descriptive Statistics of Percentiles, Minimum and Maximum Values, Standard Deviations, and Means for Each Student Demographic Predictor Variable Used in Multiple Regressions

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>P1</th>
<th>P10</th>
<th>P50</th>
<th>P90</th>
<th>P99</th>
<th>Min. value</th>
<th>Max. value</th>
<th>SD</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoC %</td>
<td>18.1</td>
<td>36.3</td>
<td>90.0</td>
<td>97.4</td>
<td>99.4</td>
<td>0.0</td>
<td>100.0</td>
<td>23.5</td>
<td>77.6</td>
</tr>
<tr>
<td>FRL %</td>
<td>5.6</td>
<td>25.1</td>
<td>79.8</td>
<td>94.5</td>
<td>97.6</td>
<td>3.9</td>
<td>100.0</td>
<td>26.7</td>
<td>69.2</td>
</tr>
<tr>
<td>EL %</td>
<td>2.5</td>
<td>7.3</td>
<td>33.6</td>
<td>62.8</td>
<td>79.5</td>
<td>1.1</td>
<td>87.0</td>
<td>20.9</td>
<td>34.3</td>
</tr>
<tr>
<td>SPED %</td>
<td>3.4</td>
<td>6.6</td>
<td>11.3</td>
<td>17.9</td>
<td>27.3</td>
<td>1.6</td>
<td>37.7</td>
<td>4.8</td>
<td>11.8</td>
</tr>
</tbody>
</table>

Note regarding abbreviations in row and column titles: “Percentile” written as “P,” and “standard deviation” written as “SD”

Figure 4.
Predicted Percent SPF Points Earned per Individual Student Demographic Variables Reflecting Models 2, 4, 6, and 8
Predicted SPF Score per Percent Free & Reduced Lunch Students

In Multiple Regression Model with Cubed Term

Predicted SPF Score per Percent English Learner Students

In Multiple Regression Model with Cubed Term
These models show a nonlinear relationship between each student demographic predictor and the percent SPF points earned. Specifically, they highlight that having greater proportions of historically marginalized students is not predicted to consistently result in the same difference in percent SPF points earned. Rather, for each of these student demographic variables there is a dramatic negative difference in predicted percent SPF points earned as a school serve these populations above the district’s lowest thresholds, and the downward trend eventually flattens out as the student demographics reach about one standard deviation below the district means.

For example, for the percent students of color variable, when holding constant all the model controls we would expect that a school with 20% students of color (representing schools in the first percentile) to have a predicted SPF score of about 78.4. Schools that serve larger proportions of students of color are predicted to earn fewer SPF points, until this negative relationship flattens out around 60% students of color, which has a predicted SPF score of 53.4 points. In this way, for every additional ten percentage points of the student of color population
in a school beginning at a base population of 20%, we would expect to see an increasingly narrowing difference in the percent of SPF points earned: schools with 30% students of color are predicted to earn about 10 fewer SPF percentage points than schools with 20% students of color; schools with 40% students of color are predicted to earn about 7 fewer SPF percentage points than schools with 30% students of color; schools with 50% students of color are predicted to earn about 4 fewer SPF percentage points than schools with 40% students of color; schools with 60% students of color are predicted to earn about 3 fewer SPF percentage points than schools with 50% students of color. After that, the predicted differences in SPF percentage points earned continues to winnow between schools with larger and larger percentages of students of color, with differences ranging from about 1 to less than 1 percent of SPF points.

The other student demographic predictors follow similar patterns. While a school with 0% FRL students is predicted to have 88.2% of SPF points earned when all the model controls are held constant, the predicted SPF points earned is considerably lower for schools with larger percentages of FRL students, with the most dramatic differences evident in schools with low percentages of FRL students compared to those with 0%, although after schools reach about 40% FRL students (predicted to earn 54.3% SPF points) the differences flatten out. Keeping constant the model controls, a school with 0% EL students is predicted to earn 71.3% of possible SPF points, but schools that serve larger proportions of ELs are predicted to initially earn dramatically fewer SPF percentage points – as schools that have 10% ELs opposed to 0% are predicted to earn about 13 fewer SPF percentage points, lowering the predicted SPF points to from 71.3% to 58.8% – until the trendline flattens at 20% ELs with a predicted SPF score of 52.0. Likewise, keeping the model controls constant a school with 0% SPED students is predicted to earn 97.5% of possible SPF points, although a school with a 5% SPED population
opposed to 0% is predicted to earn 29 fewer SPF percentage points with a predicted SPF score of 66.9%. The dramatic downward trend only continues until schools reach about 10% SPED, after which it flattens.

*Control Variables*

While this analysis focused on the student demographic variables as predictors, in a multiple regression model all variables can be interpreted as predictors or controls, and which is positioned in each role should be dictated by theory. Given this flexibility, an alternative interpretation of these data could examine the predicted percentages of SPF points earned when student demographic variables are held constant and there is variation in the percent of teachers that are Fully Qualified, the student-teacher ratio, the rate of disciplinary actions resulting in instructional loss, or the year. Although the percent of teachers that are Fully Qualified was not a statistically significant predictor of percent SPF points in any model, and student-teacher ratios were statistically significant predictors in only some models, in every model the rate of disciplinary actions resulting in instructional loss was a highly statistically significant predictor, with the year variables also being statistically significant predictors albeit with larger p-values.

This indicates that even in schools with similar student demographics, the rate at which disciplinary actions remove students from instruction is consistently a highly statistically significant predictor of percent SPF points earned, with each additional instance of such disciplinary actions per 100 students predicted to correlate with 0.2 to 0.3 lower SPF percentage points (depending on the model). Similarly, schools evaluated in the 2017-2018 academic year instead of the 2016-2017 academic year were predicted to correlate with about 3 lower SPF percentage points just for the change in year, even holding constant all the other model variables.
In addition, in 2018-2019 school SPF points were 6 points lower on average than in 2016-2017, conditional on other variables in the model.

Summary

These findings reiterate those from the previous research questions by showing that, like student demographics, school context variables extrinsic to the SPF framework do in fact correlate with SPF outcomes, making the SPF not only a reflection of student learning but school contexts, student demographics, and even the vagrancies of year. Since these data indicate that these school context variables extrinsic to the SPF had significant relationships with SPF outcomes, alternative accountability policies would do well to not only include attention to these factors but design interventions that target them.

Research Question 5: Do student demographics predict SPF outcomes?

The previous findings have shown that student demographics (and school contexts) are statistically significant predictors of percent SPF points earned, with one point positive differences in historically marginalized populations correlating with 1.5 to 8 lower SPF percentage points, depending on the model. However, a difference of only a few SPF percentage points may not be meaningful; for example, if a school earns 55% or 60% SPF points, it will still be rated in the fourth highest category of Green in the Red, Orange, Yellow, Green, and Blue SPF rating system, and it will still experience the consequent prestigious status. Because it is not the SPF percentage points themselves but the SPF outcome – whether a school is subjected to interventions or prestige – that impacts the experience of students, teachers, and families, this research question sought to understand whether the same student demographic predictor models using the same controls as the previous research question also predicted SPF outcomes broadly.
To evaluate the relationship between student demographics and SPF outcomes, I ran school-level ordinal logit regression models including cubed terms for key predictors to allow for nonlinear associations holding constant: (a) percent of teachers that are Fully Qualified, (b) student-teacher ratio, (c) number of disciplinary incidents that result in instructional loss per 100 students, and (d) dichotomous variables for academic years 2017-2018 and 2018-2019, omitting the 2016-2017 year as a reference. Each model included an individual student demographic predictor, which were: (a) percent student population that are students of color, (b) percent student population that are Special Education students, (c) percent student population that are English Learners, and (d) percent student population that receive Free and Reduced Lunch services. I then used these regressions to predict the probability of a school receiving one of three Simplified SPF designations: (a) Intervention, denoting either a Red or Orange SPF rating bracket, which warrants district intervention; (b) On Watch, denoting a Yellow SPF rating bracket, and using the term the district applies to such schools; and (c) High Performing, denoting either a Green or Blue SPF rating bracket, which are the highest ratings brackets available in the district and imply exemplary performance. The predicted margins for each academic year were used to create figures in R using ggplot (Figure 5)
Figure 5.
Predicted Probabilities of Receiving Simplified SPF Outcomes (a) Intervention, (c) On-Watch, or (c) High Performing per Student Demographic Predictor Using Models 2, 4, 6, and 8 From Research Question 4
These findings confirm those from the previous research question while adding greater nuance. In each model, there is a dramatic negative difference in the predicted probabilities of receiving a High Performing designation as schools serve proportions of historically marginalized student populations greater than 0% to about one standard deviation below the district mean for the respective population, as found in the Research Question 4. This indicates that, not only do historically marginalized student demographics predict percent SPF points earned, but having such students in proportions that deviate from the most extreme minimums appear to predict schools’ probability of receiving High Performing designations the most dramatically.

Schools with 0% students of color or Free and Reduced Lunch students have approximately a 100% predicted probability of being High Performing. At about 20% students of color and 0% FRL students, those probabilities are dramatically lower as schools serve greater proportions of these students until these populations reach around 40%, after which the changing negative trendline flattens. As schools serve greater proportions of student of color and FRL students, not only do they have steeply lower probabilities of being High Performing, but also greater probabilities of being On-Watch and Intervention until a threshold of about 50% of each student population is reached. As these student populations become greater than 0%, the upward trend of greater predicted probabilities of being On-Watch is steeper than that of being Intervention, indicating that increasing the percent of students of color or FRL students is associated with higher probabilities of being On-Watch than it does of being Intervention. The predicted probabilities of being in Intervention status are not dramatically different in the same way that the probabilities of being High Performing are after these student populations change from 0% to 40%. Instead, the predicted probabilities of being in Intervention status only begin to
show positive differences at around 40% students of color or 20% FRL students, although each
trendline begins to flatten at around 60%.

Schools with 0% ELs or SPED students have approximately 75% to 80% predicted
probability of being High Performing. As seen elsewhere, this probability is dramatically lower
for schools that serve more than 0% ELs, with downward trendlines flattening out at around 20%
for both ELs and SPED. Once a school has approximately 10% EL students, this student
population no longer predicts High Performing status more or less than On-Watch status.
Interestingly, once the percentage of ELs reaches about 60%, the predicted probability of being
High Performing is higher, meaning that schools with such large proportions of ELs have greater
predicted probabilities of being High Performing than On-Watch or Intervention. Unlike the
students of color and FRL models, the predicted probabilities of being Intervention status are
immediately higher beyond 0% ELs, although this trend flattens at around 30% ELs after which
schools with larger percentage of ELs have lower predicted probabilities of being in Intervention
status. In contrast, the predicted probability of being in Intervention status is not different until
schools have 5% or more SPED students, and this upward trend continues until 25% SPED
students after which it flattens. Also unlike the EL model, the negative difference in the
predicted probability of being High Performing as the SPED proportions are greater than 0% is
immediate, and mirrors the positive differences in the predicted probability of being On-Watch
until about a school is about 10% SPED, after which this student demographic no longer predicts
one status more than the other.

Together, these trends confirm previous findings: the percentages of historically
marginalized students predict SPF outcomes, with schools serving larger proportions of
marginalized populations having lower SPF scores and, relatedly, greater likelihood of being in
the lowest SPF category. Positive differences in the predicted probabilities of being Intervention status appear later, at about 5% for EL and SPED and 20% for students of color and FRL. Despite this nuance, a central tendency remains; given the relationship between historically marginalized student populations and SPF scores, schools that wish to earn higher SPF scores or maintain high scores have incentives to work with as small a proportion of these students as possible. As the accountability movement in education was in part rooted in a civil rights struggle to ensure public schools better served these very student populations, an accountability framework that disincentivizes schools from working with such students is a sad outcome indeed.

Summary

Findings from each of the five research questions yielded similar and mutually confirming results: In Denver Public Schools, the accountability ratings derived from the School Performance Framework not only reflect student learning but also student demographics. This results in schools that serve greater proportions of historically marginalized students – in particular students of color and students receiving Free and Reduced Lunch services – being predicted to have lower SPF ratings, potentially indicting that raced and classed students experience disparate access to educational opportunities. Furthermore, the highest- and lowest-rated schools also served markedly different types of English Learner (EL) students in markedly different ways, both in terms of these students’ stage in their trajectory toward developing English, home language, and participation in Gifted and Talented and Special Education programs, as well as in terms of what language supports these students could access. Together, these findings indicate that the SPF reflects factors extrinsic to what the framework purported to evaluate. This results in an accountability policy that implicitly disadvantages the schools that
serve more historically marginalized students while rewarding those that serve the least.

**Discussion**

The results from this study iterate a central finding: the School Performance Framework reflects and measures historically marginalized student demographics in a way that punishes the schools that serve the largest proportions of these students while failing to account for school context factors, like disciplinary rates and teacher qualifications, that also appear to drive SPF outcomes. Following this study’s commitment to a QuantCrit employment and interpretation of quantitative data, this finding is framed as a failure of both the accountability policy itself and the district that instituted it and the systemic inequalities in society that research proves causes it rather than of the historically marginalized students who are disadvantaged by the SPF or the teachers who serve them. Such an intentional reframing of the locus of responsibility for racial inequities away from racialized populations and onto institutions and the policymakers and leaders who guide them is necessary to interrupt the legacy of quantitative data being used to “obfuscate, camouflage, and even to further legitimate racist inequities” (Gillborn, Warmington & Demack, 2018; p. 160). As such, the remaining part of this chapter will focus on the ways that policymakers, researchers, and community members can use these findings as tools for pursuing greater racial and social justice rather than tools for justifying current inequities.

**Implications for Policy**

*Equity Reviews of Accountability Frameworks*

That the SPF ratings outcomes consistently reflect demographic metrics extrinsic to the accountability policy should alert district leaders and policymakers alike of the need to conduct what I call “equity reviews,” or similar data analyses as those used in this study whose purpose is to systematically review accountability data and outcomes to evaluate whether district policies
result in disproportionate harms to historically marginalized communities. The statistical methods and data used in this study are accessible to district leadership and policymakers, especially within offices of evaluation, assessment, and data management. As such, there are no logistical or methodological constraints that limit the district's ability to incorporate regular equity reviews of the outcomes of the accountability policy. It is possible to regularly incorporate equity reviews of accountability frameworks into district accountability policy. Beyond possible it is also responsible to conduct such reviews lest such policies result in the further marginalization and disadvantages of historically marginalized communities.

A primary rationale of the accountability movement was the need to ensure that schools are accountable for the outcomes of their students, especially their historically marginalized students (DeBray-Pelot & McGuinn, 2009). In the same way, the policymakers and district leaders who design and implement accountability policies should likewise hold themselves accountable for the outcomes of their work. Just as the rationale for the accountability movement holds that, when outcomes show that schools are disproportionately failing historically marginalized communities such practices are unacceptable and deserve remediation, so, too, should accountability policies and frameworks themselves be scrutinized to evaluate whether they result in disparate impacts for the very students they are intended to help. Accountability policies which result in disproportionate harm to these students should be reevaluated, revised, and dismantled as necessary. Failure to do so results in a system which discourages schools and teachers from working with historically marginalized students, as these students relate to lower accountability ratings and the negative consequences they incur.

Such a commitment to uprooting discriminatory systems is already present in the goals of the district. In 2021, the Denver Public Schools school board adopted a new governance structure
with which to orient its work and evaluate the superintendent through the use of “end goals.”

One of the “end goals” is to ensure the district is “free of oppressive systems and structures rooted in racism” (Asmar, 2022; para. 5). An accountability framework which punishes schools for working with large proportions of historically marginalized students while ignoring the disparities in school contexts, resources, and opportunities provided to those students is an example of an oppressive system rooted in racism. Whether the district uses the similar accountability framework to the one examined in this study or an alternate framework, the potential for the same disparate impacts remains. For this reason, the regularly administered equity reviews this study recommends will continue to be necessary for whichever accountability policy the district adopts. Failure to incorporate equity reviews of the district’s work and policies is both a betrayal of the district’s goals as well as a betrayal of the students and families in the district, who are subjected to externally created accountability policies, expectations, and consequences yet have no voice in how those policies are created and implemented.

Such an equity review of district accountability policies could mirror the work of this dissertation. By creating descriptive statistics of the student characteristics and school contexts in the highest- and lowest-rated schools, the district could better identify whether the accountability policy is resulting in disparate impact on certain student populations in addition to other factors – such as the dramatically different rates of disciplinary actions that result in instructional loss identified in this study – that could also have relationships with school ratings and thus deserve attention and, if necessary, amelioration as discussed in the section below “Identifying Needed Supports.” The result of a district-initiated equity review of its own work could be publicly available ratings, like those of the SPF, in which the accountability system itself is evaluated and rated. Families and teachers deserve to know whether the accountability policies used by the
district are effective and fair. Likewise, if the accountability policy results in a disproportionate amount of historically marginalized students being found in the lowest-rated schools, or worse as this study found that such student populations actually predict accountability scores, that is information the public needs as they interpret their school’s rating.

Bonilla-Silva (2006) reminds us that white supremacy is not limited to a few extremist individuals, but rather permeates the worldviews and institutions that constitute our shared social reality. The SPF is a reflection of this dynamic: although there is no part of the accountability policy, metrics, or goals that specify its design is intended to perpetuate the marginalization of historically marginalized communities, in practice (and in history) this is the outcome. Stated goals, overt acknowledgement, or even purposeful intentions are irrelevant (Leonardo, 2004), and it is likely that many if not all those who worked to craft and implement this accountability policy did so without malice. And yet, once again institutional policies and practice resulted in the same outcome, in which the marginalization of the communities already battling intergenerational marginalization not only continued but was legitimized through ostensibly ideologically-neutral metrics like test scores. Investigations like the one undertaken in this study are necessary because there is no warning label on policies that result in marginalization; there are no written statements from policymakers announcing their intent to harm historically marginalized communities as this overt intent is both likely nonexistent and certainly irrelevant when evaluating the merit and consequences of such policies.

For these reasons, conducting equity reviews of the accountability policy it is not only in line with the district’s stated goals and fair to the community, but it is an imperative in order to ensure that the district is not perpetuating these historical ills. Using a lens grounded in Critical Race Theory to analyze education policy highlights the dynamic of purportedly race-neutral
policies resulting in harms to racialized populations by asking simple questions such as, ‘Who is the policy designed by? Who does it benefit and harm? What are the outcomes?’ (Gillborn, 2005). Doing so reveals the ways that education policies such as accountability reflect, perpetuate, and legitimize the interests and worldviews of those who benefit from a white supremacists status quo at the expense of racially marked communities. Policies like the SPF – which result in schools with larger proportions of students of color being more likely to be labeled as failures and closed – perpetuate worldviews that frame marginalized students as causing their own marginalization and thus deserving of its adverse consequences while actively discouraging teachers from working with such students. This worldview is incompatible with one in which all students are capable of success and deserving of opportunities. For these reasons, this study strongly recommends that the district adopt regular equity reviews of its accountability policies, as failing to do so not only allows ineffective systems to continue but fundamentally also fails the district’s goals and responsibilities toward the communities it serves.

Evaluating Success of Accountability Framework

Such a review of the accountability framework used by the district could also identify trends of schools gaining or losing high-rated and low-rated accountability status over time, as this study found that during the study Timeframe the SPF was unsuccessful in prompting greater school success as evidenced in the downward trend in which increasing numbers of schools were given low SPF ratings while decreasing numbers of schools were given high ratings. These findings indicate that, despite the behaviorist and market logics which undergird school accountability (Trujillo & Renée, 2015), not all accountability systems will be equally successful in achieving their goals of promoting improvements in learning outcomes and school quality (Fuller & Johnson, 2001).
Incorporating regular reviews of how the accountability framework is functioning in addition to the abovementioned equity checks can help districts evaluate the efficacy of their accountability policies. Findings from this study indicate that it is possible for accountability policies to be ineffective deterrents against increasing rates of low performance. If an accountability policy is found to be ineffective in promoting the kinds of learning outcomes and quality school metrics it seeks to advance, then district leaders and policymakers would have data supporting the need to revise their policies so that the accountability frameworks they implement have better chances of succeeding in their purposes. Rather than aiming for one permanent correct system, regular review of efficacy and equity would encourage districts to engage in cycles of learning and inquiry, thus allowing them to adjust to the changing needs and contexts of the students and communities they serve.

Failure to incorporate reviews of success and efficiency of accountability frameworks risks imposing adverse consequences on students, teachers, and communities without any benefit. Like many accountability frameworks, the SPF functioned through behaviorist logics that saw negative reinforcements as mechanisms to spur desired changes in outcomes (Dworkin, 2005; Finnigan & Gross, 2007). These negative consequences included the loss of students and the funds they bring under the district’s universal choice model (Asmar, 2019a), which reduced low-rated schools’ ability to afford the teachers and programs that made them attractive destinations in the first place (Asmar, 2019b), in addition to reduced teacher pay (Asmar, 2016) and district intervention in the form of the need to complete improvement plans and possible restart or closure (Asmar, 2018; Denver Public Schools, 2018). If an accountability system results in students and teachers facing winnowing resources and enrollment, negative repute, reduced teacher pay, and possible elimination, then these adverse consequences must at
least be in the service of achieving the admirable goal of improving outcomes for students. However, if an accountability system is found to not even improve learning outcomes for students or generally promote greater school quality or school success, then these punishments only serve to adversely impact students and teachers for no purpose. As this study found that historically marginalized students are concentrated in the schools most likely to receive such punishments, it is especially imperative for districts to evaluate the effectiveness of their accountability systems lest these students and their teachers are adversely impacted for no other reason than a faulty, inefficient accountability system.

**Identifying Needed Supports**

In addition, employing similar methods as those used in this study can help districts identify school and student needs that the SPF did not address. For example, this study found that the lowest-rated schools had higher rates of discipline and lower rates of Fully Qualified teachers than the highest-rated schools. However, since neither of these metrics were measured by the SPF, as an accountability framework it was unable to identify how these disparate school contexts might have contributed to disparate achievement outcomes, thus leaving obscure information that could have helped prompt the district to offer appropriate supports and interventions. By considering the non-achievement contexts of schools that are typically disregarded by accountability frameworks, districts can help provide targeted interventions and supports that reflect actual disparities in schools, thereby ensuring that all students are provided the resources and learning environments conducive to school success.

Policymakers and district leaders could use such information to craft accountability policy interventions. For example, as this study identified disparate rates of Fully Qualified teachers in the highest- and lowest-performing schools, an appropriate intervention as a response
would be to place more Fully Qualified teachers in these schools, either by relocating them or investing in the training and incentives to ensure that the teachers already practicing at those schools can become Fully Qualified. Similarly, in response to the finding that the lowest-rated schools experience disciplinary actions, incidents, and actions resulting in instructional loss at nearly double the rate of the highest-rated schools and that rates of disciplinary actions that result in instructional loss in fact predict SPF scores, an appropriate intervention could be to provide these schools with additional support and training in restorative justice and other non-punitive social and emotional supports for students to address behavior management needs that do not result in instructional loss.

Because what is not measured is not acted upon, such a consideration for school context variables can help district leaders identify and provide the resources, services, and supports that students need but may not be receiving. Although research has long since identified the benefits of emergent bilingual students receiving bilingual education (Ramírez, 1992; Rolstad, Mahoney & Glass, 2005; Thomas & Collier, 1997), on average less than one in five English Learner students in DPS received any type of native language supports. An accountability framework which includes metrics describing whether emergent bilingual students have access to bilingual education would implicitly encourage schools to provide the resources that research has established is beneficial to these students. Furthermore, these kinds of metrics could be included in accountability frameworks that are differentiated to each school’s contexts and student needs, representing both an accountability system that understands that different student communities will have different needs as well as one that seeks to provide for those unique needs through the inclusion of differentiated metrics that encourage schools to provide them.
Another example relates to this study’s finding regarding Gifted and Talented (GT) participation between the highest- and lowest-rated schools, with Red schools identifying students for GT about half as often as Blue schools, which not only had higher rates of GT for all students but also placed English Learners in GT at five times the rate as Red schools. If dynamics such as GT participation rates were measured, then the accountability framework could use findings like the ones in this study to identify the need for providing more teacher training in cultural, linguistic, and racial biases that might prevent them from nominating students from historically marginalized backgrounds for such placement. Another response could be the inclusion of a metric evaluating the proportionate representation of historically marginalized students in GT programs in order to encourage the sorts of programmatic changes the accountability movement is premised upon. Including such metrics beyond test scores would allow accountability frameworks to identify and address these kinds of disparate school contexts and opportunities provided to students.

These are only a few examples of the ways that an accountability framework that measures school context variables can identify the contexts which distinguish high- and low-performing schools and thus provide the supports and interventions necessary to equalize the learning environments and resources students enjoy in each. As such, these findings not only highlight specific interventions that are likely necessary in Denver Public Schools, but the utility of an accountability framework that thinks outside the narrow confines of test scores to measure contextual factors and thus address root causes of achievement disparities.

Evaluating Success and Needs of Charters

Finally, this study highlighted the unique contexts of charter schools that may contribute to the understanding of the role these schools can play in improving learning outcomes for
historically marginalized students. This study found that charters were overrepresented in the lowest-rated SPF brackets (Red or Orange ratings brackets, also called Intervention Status in this study) while being less likely to enter into or remain in the highest SPF rating bracket of Blue. In addition, although on average when compared to district-run schools charters served larger proportions of historically marginalized populations of students of color, students receiving Free and Reduced Lunch services, and English Learner students, they did so in learning environments that were distinct from those of district-run schools and difficult to reconcile with providing these student populations equitable educational opportunities. For example, although on average more than 90% of EL families in charters preferred some type of language supports for their children, on average less than 4% of ELs in charters were provided with these resources. Not only were these historically marginalized students denied the language supports their families requested, but all students learned in disciplinary environments that appeared to be much harsher than those of district-run schools, with nearly double the average rates of disciplinary actions, incidents, and actions resulting in instructional loss. This dynamic suggests that, rather than rectify the shortcomings of public schools, charter schools were replicating and exacerbating some of the very same problems, such as the discipline disparities that negatively impact students of color and students in poverty (Bryan, Day-Vines, Griffin & Moore-Thomas, 2012; Skiba, Chung, Trachok, Baker, Sheya & Hughes, 2014; US Commission on Civil Rights, 2018) and the denial of adequate supports to emergent bilingual students (Redford, 2018). Without achieving improved learning outcomes as indicated by the propensity to low SPF ratings, this study suggests that in Denver Public Schools charter schools may not be the solution to public school challenges that some perceive them to be (Chubb & Moe, 2011), but instead are possibly
amplifying the inequitable practices historically marginalized students face in schools today (Kantor & Lowe, 2016).

**Implications for Researchers**

The quantitative data and methods employed by this study are not only valuable tools for district leaders and policymakers to advance more equitable educational systems for historically marginalized students, but can and should also be used by education researchers whose work advocates for these same ends. This study paid special attention to the characteristics of and services provided to students who carry the English Learner label, as these students’ frequent racialized and inherently linguistically-marked statuses have resulted in an extensive and well-documented history of these students being poorly served in public schools (Commins & Miramontes, 1989; Poza, 2016; MacSwan, 2005; San Miguel & Donato, 2010; Santa Ana, 2004; Shannon & Escamilla, 1999; Valdés, 1998). Although quantitative data and methods are not uncommon in the field of bilingual education as evidenced through assessment (Buono & Jang, 2021) and mixed methods (Hopewell, 2011) studies, this dissertation argues that expanding the use of these tools can increase the effectiveness of research advocacy in service of emergent bilingual students and families.

Currently, work which frames bilingual research and teaching as advocacy is dominated by qualitative studies (Palmer, 2018). While the roots of prioritizing qualitative data and methods when highlighting the experiences and needs of historically marginalized communities like emergent bilinguals and those whose language practices are marked is well-founded in Critical Race methods and literature (DeCuir & Dixson, 2004; Delgado, 1989), this study proposes that researchers who view their work as advocacy in service of bilingual communities would be well
served to expand those methodological approaches to include more quantitative tools in line with QuantCrit.

Understandably, there might be hesitation to use quantitative data and analysis in advocacy research, as such tools have historically been employed to legitimize the very sort of oppressive institutional policies and practices (Bonilla-Silva & Zuberi, 2008) that social justice researchers seek to dismantle. However, this study demonstrates the potential for bilingual education researcher advocates to reclaim quantitative data and analysis in service of our goals. By using such approaches, this study revealed that trends that interest the work of bilingual education researchers dedicated to promoting equitable educational experiences for bilinguals students, such as: charter schools placed English Learner students in environments designed for their success and in accordance with their family preferences at marginal rates; no ELs in the lowest-rated schools participated in Dual Language programs; that the lowest-rated schools had the largest differential between parents who wanted native language programming (averaging 40% of parents’ preferences) and ELs who received it (averaging 10% of ELs); nowhere in the district did all families who wanted native language programming for their EL students receive it (on average 39.1% of parents wanted this, but only 16.7% of ELs receive it); Spanish-speaking ELs were overrepresented in the lowest-rated schools; despite approximately one in three students being ELs, only one in 30 ELs were in GT programs; and in all schools the English Learner participation rates in Special Education programs were disproportionate to their rates in the overall student population, with this disproportionality most severe in the highest-rated schools.

Findings such as these can be used by bilingual education researchers for advocacy purposes, not only by highlighting an area in which emergent bilingual students are being
underserved and thus an area that researcher advocates should attend to, but also by providing us with quantitative data that can be easily disseminated in research publications, policy briefs, and other avenues in which we work directly with policymakers in the hopes of affecting institutional reforms. For example, the finding that Spanish-speaking ELs were overrepresented in the lowest-rated schools could inform the need for qualitative research into the raciolinguistic language ideologies of teachers regarding English-Spanish bilingualism, while the finding that there is much greater parent demand for native language programming than is currently being provided could inform policy research and advocacy to prompt districts to provide more of these services as well as schools of education to invest more in preparing the bilingual teachers necessary for these programs. Similarly, the finding that ELs participated in Special Education programs disproportionate to their rates in the overall student population could be used to establish legal standards for proving "discriminatory impact" (Haney, 2000) that bilingual education researcher advocates can then use to push district leaders, policymakers, and legislators to implement revised policies so that emergent bilingual students are better served in public schools.

Despite its problematic history, quantitative data analysis is a threshold that such decision makers use when crafting policy. Acknowledging this does not minimize the problematic tendency of quantitative data being presented as objective and value-neutral; rather, it accepts that despite being ideological in nature such types of evidence are effective in speaking to those with the power to enact the change for which we are fighting (Crawford, Demack, Gillborn & Warmington, 2019). In addition, quantitative data can be used by bilingual education researchers to inform future projects and substantiate our recommendations for increased supports and investments, both at the district as well as the teacher education context.
Implications for Teachers, Families, and Advocates

Finally, this work has particularly important implications for the students, families, and teachers who are adversely impacted by flawed accountability policies like the School Performance Framework as well as for those who consider themselves allies and advocates for such communities. Because of the potential for marginalized populations to internalize deficit ideologies (Kohli, 2014), it is necessary for research to provide counterevidence whenever possible. This research provides empirical evidence that there is something flawed in the systems used to manage and evaluate students – not something flawed in students. Absent such evidence, policies like the SPF which report that historically marginalized students are concentrated in ‘failing’ schools implicitly place the locus of responsibility of that failure on students and teachers rather than on an accountability system that is ineffective and biased, or a school system which denies them the opportunities, resources, and supports they deserve. As such, it is of little wonder why families, students, and teachers come to interpret the disparate outcomes of accountability as reflections of disparate abilities, talents, and merits. This interpretation is not only inaccurate, but deeply harmful. For this reason, this work is not only intended for policymakers and researchers but also those who are subjected to the worst outcomes of biased educational policies in hopes that offering alternative interpretations of academic disparities can deter the internalization of blame for them.

Without such data, families and teachers may come to interpret disparate accountability and achievement outcomes as reflections of personal failure. Even if this is not the case, families may erroneously believe that, if they can only find an alternative school such as through the allure of charters, their students will have greater opportunities and success. Sadly, the results from the regression analyses in this study suggest otherwise, as findings indicated that the
proportion of students of color in a school was a statistically significant predictor of both accountability scores as well as outcomes. This means that, if schools are subjected to biased accountability frameworks or school systems do not provide equitable resources and opportunities to students of color, there is no “escaping” these low ratings and disparate outcomes; rather, they follow students as the low ratings are related to student demographics and not necessarily student learning or school quality. Since student demographics predict accountability scores, if those students change schools the low accountability scores are predicted to follow them. Such information is important for families and teachers to have, not only as it displaces the blame for low accountability scores from them personally but also because it clarifies that simply changing schools is unlikely to result in improvements as the accountability system, and possibly the district mechanisms for allocating resources and opportunities, are the cause of the disparate outcomes, not the students or the schools.

Using methods and data like those employed in this study can help to directly counter such deficit interpretations of academic disparities by highlighting the role that non-student, non-teacher, and non-family factors play in producing disparate outcomes such as the overrepresentation of historically marginalized students in low-rated schools. Just as Critical Race scholars use counterstories to contest dominant deficit narratives (Ladson-Billings, 2013a), so too can quantitative research that highlights the institutional mechanisms by which historically marginalized communities are further disenfranchised be taken as a counternarrative, as such data dispels interpretations which place the blame for academic disparities on those communities themselves. Teachers, families, and students who are subjected to deficit narratives that attribute the responsibility for academic disparities to them personally deserve to have access to counter evidence that more accurately ascribes responsibility to the policymakers and district leaders.
who craft and enact accountability frameworks and educational policies that lead to biased outcomes and inadequate attention to students’ needs.

Further, these data and methods are accessible to a wide range of audiences. As someone with only a limited background in statistical methods, the research design of this study by necessity reflects an intuitive approach to quantitative data and analysis that I have found to be accessible to the teachers, advocates, and families with whom I share this work. As such, for teachers, families, students, and their allies, the approach used in this dissertation can offer a means to communicate about and understand educational disparities that is accessible to policymakers and community members alike, thus not only offering counternarratives to combat the internalization of deficit views but also tools to advocate for educational policies and practices that better serve historically marginalized students.

Limitations

This is not to say the study is without limitations. The focus of the study, the School Performance Framework, was disbanded in 2020 and replaced by the accountability framework developed and used by the Colorado Department of Education, also called the School Performance Framework (Denver Public Schools (n.d. - d), making the issues and shortcomings explored here without a current referent in the district. However, because of the centrality of racism to US history and institutions (Ladson-Billings, 2013a), similar investigations of other iterations of accountability policies will still be needed, as the disparate impacts of accountability frameworks that disadvantage historically marginalized communities is not isolated to Denver (Glynn & Waldeck, 2013; Harris, 2007; Lakin & Young, 2013; Martinez-Garcia, LaPrairie & Slate, 2011; McNeil, Coppola, Radigan & Vasquez Heilig, 2008; Menken, 2006; Reyes & Garcia, 2014; Tsang, Katz & Stack, 2008; Vasquez Heilig & Darling-Hammond, 2008; Wu,
2013). Nonetheless, that this study describes the specific nature and outcomes of an accountability policy that is no longer used represents a limitation of the utility of the findings, although the implications for policymakers, researchers, and community members to similarly employ quantitative data and methods remain, as does the need for future iterations of accountability to be reviewed for equity and efficacy not only in Denver but in any district employing similar accountability frameworks.

In addition, because the multiple regressions in this study treated student demographic categories as discrete rather than intersectional, this study may perpetuate inaccurate representation of student identities that compromises the utility and accuracy of the findings (Covarrubias & Vélez, 2013; Covarrubias, Nava, Lara, Burciaga, Vélez & Solórzano, 2018). Future studies employing similar methods for similar purposes would be well served to expand the methodological framework in order to produce more nuanced findings and more accurately represent the intersectional identities of the historically marginalized communities at the heart of this study.

Conclusion

This study used QuantCrit and Critical Race frameworks to examine the student demographic, school context, and English Learner characteristics and services that previous research suggested impact the learning opportunities and outcomes of students but that were extrinsic to the accountability framework used in Denver Public Schools. The central finding of this study confirmed the need for this analysis, as these factors were all reflected in accountability outcomes yet not officially measured by the accountability framework. This finding indicates that the SPF used by Denver Public Schools was not solely a measure of student learning or school quality, but also student demographics, school contexts, and English
Learner characteristics and services. Yet, without actually measuring these factors, the accountability framework was unable to identify and respond to how they appear to relate to the disparate learning outcomes that the SPF did measure. The disconnect between the learning outcomes the SPF purported to measure and these extrinsic factors which it in reality reflected resulted in an accountability framework that had limited ability to identify the needs of low-performing schools and thereby provide needed interventions and supports, which likely accounts for the finding that more and more schools become low-performing over time despite the intention of the accountability framework to have the opposite effect.

This study recognizes that the accountability movement has roots in the struggle of historically marginalized communities to create more equitable learning environments for their students. Yet the way accountability was manifested in Denver Public Schools appears to have had the opposite effect, penalizing schools and teachers for working with larger proportions of these students and offering solutions in the form of charters which further exacerbated the inequitable environments and supports these students received. As such, this study highlights the need for accountability policy to be more intentional in its design and implementation, with a greater focus on evaluating non-test metrics of school needs and contexts in order to provide the supports necessary to equalize the learning environments and opportunities between the lowest- and highest-rated schools. Other non-test metrics like student demographics must also be measured to ensure accountability systems are not reproducing the disenfranchisement of historically marginalized communities. This study suggests that equity checks be incorporated into any accountability policy to ensure that adverse impacts are not disproportionately felt by historically marginalized students, in addition to the publication of the outcomes of these checks so students and families can evaluate both the efficacy and fairness of the accountability results.
Perhaps more importantly, this study used quantitative data to highlight the disparate school contexts, services, and types of students in the highest- and lowest-rated schools as a means of providing a counternarrative to the deficit view which would ascribe disparate learning outcomes to student and teacher failure. The finding that the accountability policy employed by the district reflected student demographic and school contexts metrics extrinsic to the framework is a valuable counternarrative to the deficit ideologies which would ascribe responsibility of educational disenfranchisement to the communities that suffer them rather than district leaders and policymakers who allocate resources and opportunities. The research methods and data employed here are not only fruitful means of producing such counternarratives, but they can also be useful tools for policymakers, bilingual education researchers and advocates, and community members and allies who likewise seek to identify the mechanisms by which educational policy reproduces and legitimates marginalization. Doing so can help us explore how educational policies can then be revised and thus converted into a means of equitably serving and empowering the students of color, students in poverty, and especially emergent bilingual students for whom I hope this study has been of service.

Although I will never see, experience, or understand the world like they do, I have been witness to the casual, chronic, and systemic abuses that my friends and family have endured in public schools. Because of this, I hope this work is successful not only in exposing the ways that accountability policy results in marginalization, but also in aiding the pursuit of better educational systems that treat all children with the love and humanity they deserve. This investigation into the disparate outcomes of accountability policy strives to highlight the places where current policy fails to serve the raced, classed, and linguistically marked students in
Denver. In doing so, I hope this project serves the work of all those who strive toward creating a better, more equitable system for daughter, nieces, nephews, and all the students like them.
References


Asmar, M. (2020a, June 10). *Black students in Denver are much more likely to be ticketed or arrested at school*. Chalkbeat. https://co.chalkbeat.org/2020/6/10/21287249/black-students-denver-more-likely-ticketed-arrested


Burman, E., Greenstein, A., Bragg, J., Hanley, T., Kalambouka, A., Lupton, R., ... & Winter, L.


Consent Decree of the U.S. District Court (2012). *Consent Decree of the U.S. District Court: Denver Public Schools; English Language Acquisition Program*. Denver Public Schools.


https://nvdatabase.swarthmore.edu/content/east-los-angeles-students-walkout-educational-reform-east-la-blowouts-1968


Law Review, 87(8), 2411-2441.

https://drive.google.com/file/d/1fXoljVjQShaji8kAsljcYb0DHMW2jGHve/view

https://portfolio.dpsk12.org/school-performance-compact/


Dorn, S., & Ydesen, C. (2014). Towards a comparative and international history of school testing


Ladson-Billings, G. (2013a). Critical race theory: What it is not! In M. Lynn, & A. D. Dixson,
151


Losen, D. J., & Martinez, P. (2020). *Lost opportunities: How disparate school discipline continues to drive differences in the opportunity to learn*. Palo Alto, CA/Los Angeles, CA: Learning Policy Institute; Center for Civil Rights Remedies at the Civil Rights Project, UCLA.


Educational evaluation and policy analysis, 17(3), 305-322.


performance in two urban school districts. *Educational Policy, 17*(5), 558-585.


Turkan, S., & Buzick, H. M. (2016). Complexities and issues to consider in the evaluation of

US Commission on Civil Rights. (2018). *Public education funding inequity: In an era of
increasing concentration of poverty and resegregation.*

*Educational Researcher, 27*(6), 4-18.

Albany: State University of New York Press.


investigation of Society’s educational debts due to racism and sexism in chemistry

and learning of urban minority students in a high-stakes testing context. *Educational

Vasquez Heilig, J. (2011). As good as advertised? tracking urban student progress through high

Vasquez Heilig, J., Young, M., & Williams, A. (2012). At-risk student averse: Risk management


# Appendix A

## Appendix Table 1.
*Data Sources, Datasets, Data Types, and Data Uses in Dissertation*

<table>
<thead>
<tr>
<th>Source of Dataset</th>
<th>Description of Dataset</th>
<th>Variables within Dataset</th>
<th>Use in Dissertation</th>
<th>Type of Data</th>
</tr>
</thead>
</table>
| **CDE Education Statistics** | Suspension/Expulsion Statistics – Discipline by Race/Ethnicity | • “ Destruction of school property”  
• “Detrimental behavior”  
• “Disobedient”  
• “Code of conduct”  
• “Classroom suspension”  
• “In-school suspension”  
• “Out of school suspension”  
• “Other action,”  
• “Expulsion”  
• “Referral to law enforcement” | • Counts used to calculate rate of disciplinary actions and incidents per 100 students | • Counts |
| **CDE Education Statistics** | Pupil Membership – Race/Ethnicity | • “ Students of Color”  
• “PreK-12 Total Enrollment” | • Counts used to calculate percentage of students of color out of total enrollment | • Counts |
| **CDE Education Statistics** | Pupil Membership by Service Types | • “Gifted and Talented”  
• “Special Education”  
• “English Learner”  
• “Free and Reduced Lunch” | • Counts used to calculate percentages of students per service type | • Counts |
| **CDE Education Statistics** | School/District Staff Statistics | • “Student Teacher Ratios” | • No changes made | • Percent (rate) |
| **DPS SPF Reports** | All Schools SPF Indicator Summary Report | • “SPF Rating”  
• “SPF Earned Points %” | • No changes made | • Percent  
• Categorical variables |
| **CD July Report** | 9VC5 | • “Redesignated English Learners Count”  
• “Re-Entered English Learners Count”  
• “Exited English Learner Count”  
• “ELA Program Type”  
• “School Type” [district-run or charter] | • Counts used to calculate rate of Redesignations, Exits, and Re-Entries per total English Learner population | • Counts by language status  
• Categorical variables |
| **CD July Report** | 9VC11 | • “Teacher total”  
• “Fully qualified teacher counts” | • Teacher counts used to calculate percentage of fully qualified teachers | • Counts |
| CD October Report | 9VC23 | • “Gifted and Talented – English Learner %”  
• “Gifted and Talented – Never English Learner %”  
• “Gifted and Talented – Exited English Learner %”  
• “Gifted and Talented – Redesignated English Learner %” | • No changes made | • Percent |
|------------------|-------|-------------------------------------------------|----------------|----------|
| CD July Report   | 9VA2  | • “WIDA Access Scores”  
• “English Learner/Provisional Total”  
• “English Learner/Provisional by Language Status”  
• “English Learners in Special Education Total”  
• “N of all ELs” per PPF category  
• “N of all ELs” per program setting | • ACCESS scores 1-2, 3-4, and 5-6 were combined to make the respective “Beginner,” “Intermediate,” and “Advanced” Level counts  
• Those counts were used to calculate the percentage of ELs per Level  
• Spanish-speaking ELs were totaled; that total was used to calculate percentage of Spanish-speaking ELs  
• Counts of ELs per PPF and program settings were used to calculate the percentages of students in each category | • Counts |
Appendix B

Appendix Table 2.
Means of Student Demographics, English Learner Characteristics, English Learner Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Year 2016-2017

<table>
<thead>
<tr>
<th>School Characteristics</th>
<th>2016-2017 Academic Year</th>
<th>District Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Red</td>
<td>Orange</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>%</td>
<td>4.7%</td>
<td>7.4%</td>
</tr>
<tr>
<td><strong>Student Demographics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students of Color %</td>
<td>89.7</td>
<td>84.5</td>
</tr>
<tr>
<td>Free and Reduced Lunch %</td>
<td>83.8</td>
<td>77.5</td>
</tr>
<tr>
<td>Special Education %</td>
<td>15.0</td>
<td>14.8</td>
</tr>
<tr>
<td>English Learner %</td>
<td>28.6</td>
<td>40.6</td>
</tr>
<tr>
<td>Gifted and Talented %</td>
<td>10.6</td>
<td>11.3</td>
</tr>
<tr>
<td><strong>English Learner Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special Education as English Learners %</td>
<td>38.0</td>
<td>42.2</td>
</tr>
<tr>
<td>Spanish-Speaking English Learner %</td>
<td>81.0</td>
<td>80.7</td>
</tr>
<tr>
<td>English Learners in Gifted and Talented %</td>
<td>2.3</td>
<td>3.9</td>
</tr>
<tr>
<td>Beginning Level English Learner %</td>
<td>20.3</td>
<td>23.6</td>
</tr>
<tr>
<td>Intermediate Level English Learner %</td>
<td>76.1</td>
<td>70.5</td>
</tr>
<tr>
<td>Advanced Level English Learner %</td>
<td>3.5</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>English Learner Services</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redesignation %</td>
<td>3.8</td>
<td>6.1</td>
</tr>
<tr>
<td>Exit %</td>
<td>6.0</td>
<td>6.3</td>
</tr>
<tr>
<td>Re-Entry %</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>Parent Preference 1 % (bilingual ed)</td>
<td>33.0</td>
<td>47.0</td>
</tr>
<tr>
<td>Parent Preference 2 % (whatever is at school)</td>
<td>57.6</td>
<td>44.8</td>
</tr>
<tr>
<td>Parent Preference 3 % (nothing)</td>
<td>6.0</td>
<td>10.1</td>
</tr>
<tr>
<td>Mainstream %</td>
<td>24.8</td>
<td>47.1</td>
</tr>
<tr>
<td>ELA - English %</td>
<td>65.9</td>
<td>36.8</td>
</tr>
<tr>
<td>ELA – Spanish (ELAS) %</td>
<td>9.3</td>
<td>16.1</td>
</tr>
<tr>
<td>Dual Language (DL) %</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Native Language (ELAS+DL) %</td>
<td>9.3</td>
<td>16.1</td>
</tr>
<tr>
<td><strong>School Contexts</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Enrollment</td>
<td>287.4</td>
<td>382.2</td>
</tr>
<tr>
<td>Student-Teacher Ratio</td>
<td>15.5</td>
<td>15.5</td>
</tr>
<tr>
<td>Fully Qualified Teacher %</td>
<td>70.1</td>
<td>72.8</td>
</tr>
<tr>
<td>Disciplinary Actions per 100 Students</td>
<td>34.1</td>
<td>19.1</td>
</tr>
<tr>
<td>Disciplinary Incidents per 100 Students</td>
<td>29.3</td>
<td>11.0</td>
</tr>
</tbody>
</table>
Disciplinary Actions Resulting in Instructional Loss per 100 Students

<table>
<thead>
<tr>
<th>Charter School %</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.0</td>
</tr>
<tr>
<td>42.9</td>
</tr>
<tr>
<td>20.8</td>
</tr>
<tr>
<td>27.1</td>
</tr>
<tr>
<td>40.0</td>
</tr>
<tr>
<td>29.0</td>
</tr>
</tbody>
</table>

### Appendix Table 3.
*Means of Student Demographics, English Learner Characteristics, English Learner Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Year 2017-2018*

<table>
<thead>
<tr>
<th>School Characteristics</th>
<th>Red</th>
<th>Orange</th>
<th>Yellow</th>
<th>Green</th>
<th>Blue</th>
<th>District Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>17</td>
<td>20</td>
<td>71</td>
<td>74</td>
<td>12</td>
<td>194</td>
</tr>
<tr>
<td>%</td>
<td>8.8%</td>
<td>10.3%</td>
<td>36.6%</td>
<td>38.1%</td>
<td>6.2%</td>
<td>100%</td>
</tr>
</tbody>
</table>

#### Student Demographics
- Students of Color %: 86.4, 88.1, 77.1, 79.0, 46.5, 80.6
- Free and Reduced Lunch %: 75.8, 79.9, 70.7, 70.8, 34.6, 71.0
- Special Education %: 15.6, 13.9, 11.9, 11.5, 8.5, 11.4
- English Learner %: 40.0, 37.6, 32.4, 39.3, 19.4, 37.1
- Gifted and Talented %: 13.6, 16.4, 14.7, 13.5, 19.2, 14.8

#### English Learner Characteristics
- Special Education as English Learners %: 42.3, 43.4, 35.2, 40.9, 19.9, 42.8
- Spanish-Speaking English Learner %: 88.1, 85.7, 80.7, 78.0, 52.2, 82.5
- English Learners in Gifted and Talented %: 2.7, 1.9, 2.4, 2.2, 13.5, 2.5
- Beginning Level English Learner %: 20.2, 25.1, 23.1, 23.0, 13.7, 17.3
- Intermediate Level English Learner %: 76.1, 70.5, 71.5, 69.9, 67.7, 78.3
- Advanced Level English Learner %: 3.6, 4.3, 5.4, 7.1, 18.6, 4.4

#### English Learner Services
- Redesignation %: 9.7, 9.9, 17.6, 10.0, 15.7, 15.9
- Exit %: 4.1, 4.3, 5.5, 5.2, 9.9, 7.5
- Re-Entry %: 0.4, 0.7, 0.2, 0.3, 0.9, 0.8
- Parent Preference 1 %: 39.6, 37.7, 39.2, 41.7, 18.2, 38.3
- Parent Preference 2 %: 47.5, 54.6, 52.4, 53.3, 72.0, 50.1
- Parent Preference 3 %: 12.0, 7.1, 7.7, 4.5, 9.7, 11.3
- Mainstream %: 38.4, 40.6, 22.7, 27.7, 29.7, 33.4
- ELA - English %: 52.3, 48.4, 59.9, 50.7, 66.2, 58.1
- ELA – Spanish (ELAS) %: 9.3, 11.0, 15.9, 17.3, 4.2, 6.4
- Dual Language (DL) %: 0.0, 0.0, 1.4, 4.2, 0.0, 2.1
- Native Language (ELAS+DL) %: 9.3, 11.0, 17.4, 21.6, 4.2, 8.5

#### School Contexts
- Total Enrollment: 299.6, 496.7, 454.8, 434.8, 470.4, 612.1
- Student-Teacher Ratio: 14.5, 13.8, 14.5, 14.9, 16.6, 14.7

164
Appendix Table 4.
Means of Student Demographics, English Learner Characteristics, English Learner Outcomes and Programs, and School Contexts Across SPF Ratings Brackets for Academic Year 2018-2019

<table>
<thead>
<tr>
<th>School Characteristics</th>
<th>2018-2019 Academic Year</th>
<th>2018-2019 Academic Year</th>
<th>District Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Red</td>
<td>Orange</td>
<td>Yellow</td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>23</td>
<td>69</td>
</tr>
<tr>
<td>%</td>
<td>12.6%</td>
<td>12.0%</td>
<td>36.1%</td>
</tr>
<tr>
<td>Student Demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students of Color %</td>
<td>87.5</td>
<td>85.6</td>
<td>77.6</td>
</tr>
<tr>
<td>Free and Reduced Lunch</td>
<td>78.5</td>
<td>73.2</td>
<td>70.4</td>
</tr>
<tr>
<td>Special Education %</td>
<td>15.5</td>
<td>14.6</td>
<td>12.5</td>
</tr>
<tr>
<td>English Learner %</td>
<td>37.4</td>
<td>39.1</td>
<td>31.8</td>
</tr>
<tr>
<td>Gifted and Talented %</td>
<td>9.2</td>
<td>8.4</td>
<td>9.8</td>
</tr>
<tr>
<td>English Learner Services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special Education as English Learners %</td>
<td>42.1</td>
<td>45.4</td>
<td>33.4</td>
</tr>
<tr>
<td>Spanish-Speaking English Learner %</td>
<td>88.4</td>
<td>86.5</td>
<td>77.9</td>
</tr>
<tr>
<td>English Learners in Gifted and Talented %</td>
<td>1.3</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td>Beginning Level English Learner %</td>
<td>28.4</td>
<td>27.3</td>
<td>26.0</td>
</tr>
<tr>
<td>Intermediate Level English Learner %</td>
<td>68.0</td>
<td>66.7</td>
<td>68.0</td>
</tr>
<tr>
<td>Advanced Level English Learner %</td>
<td>3.7</td>
<td>6.1</td>
<td>6.0</td>
</tr>
<tr>
<td>English Learner Services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redesignation %</td>
<td>13.5</td>
<td>11.7</td>
<td>17.5</td>
</tr>
<tr>
<td>Exit %</td>
<td>8.8</td>
<td>8.5</td>
<td>7.4</td>
</tr>
<tr>
<td>Re-Entry %</td>
<td>1.0</td>
<td>2.7</td>
<td>1.9</td>
</tr>
<tr>
<td>Parent Preference 1 %</td>
<td>42.8</td>
<td>44.0</td>
<td>38.0</td>
</tr>
<tr>
<td>Parent Preference 2 %</td>
<td>50.2</td>
<td>47.5</td>
<td>55.8</td>
</tr>
<tr>
<td>Parent Preference 3 %</td>
<td>8.3</td>
<td>8.9</td>
<td>7.8</td>
</tr>
<tr>
<td>Mainstream %</td>
<td>5.9</td>
<td>25.0</td>
<td>7.3</td>
</tr>
<tr>
<td>ELA - English %</td>
<td>82.4</td>
<td>50.9</td>
<td>77.1</td>
</tr>
<tr>
<td>ELA – Spanish (ELAS) %</td>
<td>11.7</td>
<td>15.0</td>
<td>15.6</td>
</tr>
<tr>
<td>Dual Language (DL) %</td>
<td>0.0</td>
<td>9.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Native Language (ELAS+DL) %</td>
<td>11.7</td>
<td>24.1</td>
<td>15.6</td>
</tr>
</tbody>
</table>
### School Contexts

<table>
<thead>
<tr>
<th></th>
<th>333.9</th>
<th>380.9</th>
<th>500.3</th>
<th>437.0</th>
<th>455.5</th>
<th>616.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Enrollment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student-Teacher Ratio</td>
<td>13.7</td>
<td>14.7</td>
<td>14.5</td>
<td>14.8</td>
<td>16.0</td>
<td>14.7</td>
</tr>
<tr>
<td>Fully Qualified Teacher %</td>
<td>75.0</td>
<td>74.7</td>
<td>82.8</td>
<td>83.4</td>
<td>83.9</td>
<td>80.6</td>
</tr>
<tr>
<td>Disciplinary Actions per 100 Students</td>
<td>24.3</td>
<td>14.5</td>
<td>17.7</td>
<td>8.1</td>
<td>10.7</td>
<td>16.4</td>
</tr>
<tr>
<td>Disciplinary Incidents per 100 Students</td>
<td>16.0</td>
<td>8.6</td>
<td>12.4</td>
<td>5.9</td>
<td>6.6</td>
<td>10.4</td>
</tr>
<tr>
<td>Disciplinary Actions Resulting in Instructional Loss per 100 Students</td>
<td>8.8</td>
<td>5.6</td>
<td>6.3</td>
<td>2.6</td>
<td>2.5</td>
<td>5.4</td>
</tr>
<tr>
<td>Charter School %</td>
<td>52.2</td>
<td>45.5</td>
<td>17.7</td>
<td>31.6</td>
<td>26.7</td>
<td>30.3</td>
</tr>
</tbody>
</table>