

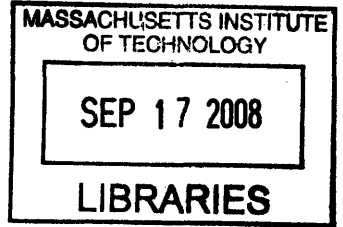
The Minority Achievement Gap in a Suburban School District

by

Lincoln Jamond Chandler

S.M., Operations Research
Massachusetts Institute of Technology, 2005

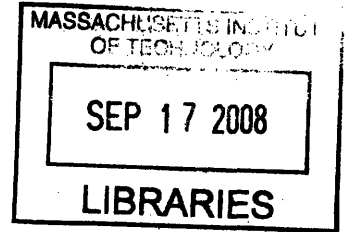
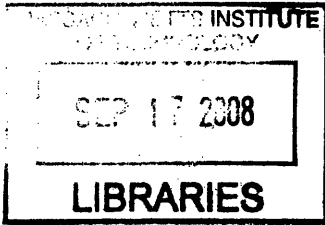
B.S., Computer Information Sciences
Florida Agricultural & Mechanical University, 1999



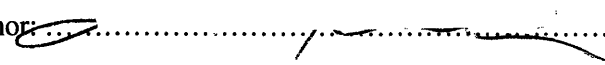
Submitted to the Sloan School of Management
in Partial Fulfillment of the Requirements for the Degree of

Ph.D. in Operations Research
at the
Massachusetts Institute of Technology

June 2008



© 2008 Massachusetts Institute of Technology
All rights reserved

Signature of Author: 
Operations Research Center
April 31, 2008

Certified by:
Arnold I. Barnett
George Eastman Professor of Management Science
Thesis Supervisor

Accepted by:
Cynthia Barnhart
Professor, Department of Civil and Environmental Engineering
Co - Director, Operations Research Center

ARCHIVES

ARCHIVES

The Minority Achievement Gap in a Suburban School District

by

Lincoln Jamond Chandler

Submitted to the Sloan School of Management
on April 31, 2008 in Partial Fulfillment of the
Requirements for the Degree of PhD in
Operations Research

ABSTRACT

For many decades, the American educational system has yielded significant differences in achievement among students in different racial groups, a phenomenon commonly known as the “Achievement Gap”. Despite the volume of research devoted to studying achievement gaps, school administrators faced with the challenge of reducing these gaps have had limited success. There are a number of factors, regarding the individual, the school, and the setting, which can contribute to achievement gaps, but in a particular community, the prevalence of such factors, and their individual contribution to the gap, is unclear.

In this dissertation, we employ a variety of statistical methods that provide a bridge between large-scale studies of achievement gaps and the analyses necessary to address the needs of a single community. First, we establish a collection of metrics designed to measure relative and absolute differences in achievement, for groups of arbitrary size and distribution. Using data from a middle-class, racially integrated school district, we employ these metrics to measure the magnitude of the achievement gap for individual students from grades three through eight. We also assess the potential role of previously-identified correlates of low achievement, such as poverty and student mobility. Last, we evaluate the potential impact of strategies for narrowing the gap.

Thesis Supervisor: Arnold I. Barnett
Title: George Eastman Professor of Management Science

ACKNOWLEDGEMENTS

*Foremost, I give thanks to my family: **Raymond D., Audrey P., and Simon J. Chandler**; in all that I do, your love and your support are my foundation. I submit this thesis in celebration and appreciation of you.*

*To my committee: **Arnie**, I can not adequately express my gratitude for your friendship and guidance. At a point when I seriously considered leaving grad school, you stepped in to help me find a topic that would fit. Three years later, I'm not sure how much of this journey went according to plan, but I am a better thinker, writer, and person for it. Sincere thanks to **Ron Ferguson** and **Amedeo Odoni** for their participation in the development of my topic and their helpful comments in framing the work.*

*To Dr. Collins, Dr. Anderson, Oak Park #97, and the Oak Park community: Thank you so much for inviting me into your community and providing the foundation for my research. Your collective efforts toward closing achievement gaps are inspiring and I hope that my analyses were in some way helpful to your ongoing work. I thoroughly enjoyed working with **Mark Pickus**; as my primary contact and tremendously helpful in helping me navigate and understand the district data. Many thanks also to **Dr. Oliver Pergams** for establishing my relationship with Oak Park; his vigilance as a parent first, and as an academic second, made this work possible.*

*To my department, the Operations Research Center: I could not have asked for a better group of friends and colleagues. **Paulette** and **Laura**, I'm not sure how many deadlines I've missed over the years, but thank you for putting up with me for so long, in spite of myself. **Jim Orlin** and **Dick Larson**, I thank you for helping me find my way in the early years. **Dimitris Bertsimas** and **Cindy Barnhart**, I thank you for your patience and counsel in the later years. I think my ORC tenure is second only to faculty and staff, so I can't possibly rattle off the names of all of my ORC buddies, but you guys are the best. I would say goodbye on behalf of the entering **2001 cohort**, but I believe Eric Z. has just returned, so never mind.*

*Thanks to my family of educators: These folks were instrumental in introducing me to the education landscape and equipping me with the background and encouragement to dive in: **Robbin Chapman, Shani Daily, Amon Millner, and Karl Reid**, all from MIT; **Kay Merseth** and the **HGSE** crew of 2005; **Scott Morgan, Sionainn Marcoux**; and my fellow **Education Pioneers**; and **Mike Webb, Joel Vargas**, and the great folks at **Jobs for the Future**. Thanks for helping a "math guy" feel welcome.*

*To my extended MIT and Boston family: My MIT odyssey began in the Fall of 1997 in Tallahassee, FL. I had just finished interviewing for a summer internship and was waiting for a bus to campus when I met **Dean Staton** of the **Graduate Students Office**. She told me about the **MIT Summer Research Program**, and 11 years later, here we are. Special thanks to **Ike Colbert** for his advice, commitment, and advocacy. I thank the members of **BGSA** and the **BSU** for providing a cultural home away from my **FAMU** roots, and my friends from **BGM** for fellowship and friendship. And in particular, I must thank that special group of people who strive to reach the **ACME** of their respective disciplines through dedication to accountability and getting goals accomplished.*

*Lastly, I dedicate this thesis to two very special women in my life. **Mrs. Pauline Person**, my grandmother, is a hero and an inspiration in my life. We lived in the same house until I was 10*

years old and, throughout my life, no matter where I was, she would check up on me and encourage me to “get my lesson.” She passed away September 15, 2007, but her spirit and memory inspires and encourages me still. **Miss Jordyn Simone Chandler** is my niece, born February 5 of this year. In her eyes, I see unbounded possibilities for her future and a hunger for knowledge and understanding of the world around her. I present this work in the hope that little Jordyn maintains that hunger throughout her life, and in partial fulfillment of my desire to do what I can to nurture our youth and address any and all obstacles that block the path to knowledge.

TABLE OF CONTENTS

1. Introduction	11
1.1. Background	11
1.2. Objective	12
1.3. Findings	13
1.4. Overview of Chapters	16
2. Background	19
2.1. Historical Perspectives on Race and Achievement	19
2.1.1. The Coleman Report (1966)	19
2.1.2. The Persistence of Achievement Gaps	20
2.2. Key Issues Regarding Achievement Gaps	22
2.2.1. External Effects versus School Effects	22
2.2.2. Achievement Gaps among the Middle-Class	28
2.2.3. Measurement and Reporting	34
2.3 Our Contributions	36
3. Testing and Measurement	39
3.1. Measuring Achievement Gaps	39
3.1.1. Average Score	40
3.1.2. Performance Levels	40
3.1.3. Standardized scores	45
3.2. Alternative Measures of Achievement Gaps	52
3.2.1. Relative Achievement Gaps	53
3.2.2. Absolute Achievement Gaps	61
3.3. Absolute Metrics vs. Relative Metrics	64
3.4. Summary	65
4. The 8th Grade Mathematics Gap in Oak Park	67
4.1 District Overview	68
4.2 The ISAT Mathematics Examination	69
4.2.1. Scoring the ISAT exam	70

4.2.2. Scale Recalibration	71
4.3. Grade 8 Test Performance (2005)	73
4.3.1. Performance Summary	73
4.3.2. Measuring the Test Performance Gap	74
4.4 Grade 8 Classroom Performance (2005)	76
4.4.1. Overview	77
4.4.2. Methods	78
4.4.3. Findings	80
4.5. Summary	81
5. Evolution of the Achievement Gap	83
5.1. The Achievement Gap across Cohorts	83
5.1.1. Comparing the Cohorts	83
5.1.2. Testing Similarity across Cohorts	85
5.2 Changes in Group Achievement over Time	88
5.2.1. Comparing Test Gaps across Grade Levels	88
5.2.2. Pre- ISAT Achievement Gaps	89
5.3. Changes in Student Progress over Time	92
5.3.1. Group Achievement vs. Student Progress	92
5.3.2. Transition Frequencies	93
5.3.3. Statistical Significance	97
5.3.4. A Closer Look at Math3 – Math8 Progress	99
5.4. Summary	102
6. Correlates of Achievement in Oak Park	105
6.1. Student Mobility	105
6.1.1. Overview	105
6.1.2. Mobility Trends in Oak Park	106
6.2. Family Income	109
6.2.1. Overview	109
6.2.2. Low-Income Status	109
6.2.3. Census Income Data	112

6.3 Gender	116
6.3.1. Overview	116
6.3.2 Gender and Achievement across Ethnic Groups	117
6.3.3 Gender and Achievement within Ethnic Groups	120
6.4. Chapter Summary	123
7. Potential Levers for Reducing Achievement Gaps	127
7.1. Academic Transition Rates	127
7.1.1. Overview	127
7.1.2. Methods	129
7.1.3. Findings	131
7.2 Reading Comprehension	135
7.2.1. Overview	135
7.2.2. Methods	136
7.2.3. Findings - Minority Males	139
7.2.4. Findings for other groups	141
7.2.5. Potential Impact on Math Achievement	143
7.3. Middle School Honors Enrollments	145
7.3.1. Middle School Math in Oak Park	145
7.3.2. Course Enrollments and the Eighth Grade Achievement Gap	147
7.3.3. Grade 8 Placement Rates.	153
7.3.4. Estimated Impact of Equal Honors Placement	159
7.4. Teacher Effectiveness	162
7.4.1. Background	162
7.4.2. Methods	163
7.4.3. Results	168
7.4.4. Other Comments	173
7.5. Chapter Summary	174
8. Caveats	177

9. Conclusions and Future Work	181
9.1. Measurement	181
9.2. Achievement Gap Dynamics	182
9.3. Potential Correlates of Achievement Gaps	183
9.3.1. Low-Income and Student Mobility	184
9.3.2. Achievement and Gender	184
9.3.3. Achievement and Reading Comprehension	185
9.3.4. Achievement and Honors Placement	186
9.3.5. Achievement and Teachers	187
9.4. Comments	187
10. References	189

1. Introduction

1.1. BACKGROUND

For many decades, the American educational system has yielded significant differences in achievement among students in different racial groups (Coleman et al., 1966; Harris and Herrington, 2006). A common name for the phenomenon is the “Achievement Gap”. Studies of the Achievement Gap typically concern performance gaps between “Black” and “White” racial groups; with respect to other racial groups, Asian students tend to align with White students, and Hispanic and Native American students tend to align with Black students.

Long recognized as an issue of national importance, the Achievement Gap has received increased scrutiny in recent years. With the passage of a federal measure known as the No Child Left Behind Act of 2001 (“NCLB”), school districts have been charged with the task of bringing all students to academic proficiency. In keeping with NCLB, school leaders are now required to report achievement test data by race¹. The added transparency of student outcomes make explicit the notion that school leaders and superintendents are to held accountable for monitoring and improving the academic performance and progress of all groups of students.

Despite the volume of research devoted to studying achievement gaps, school administrators faced with the challenge of reducing these gaps have had limited success. There are a number of factors regarding the individual, the school, and the environment that influence to achievement gaps, but at the community level, the prevalence of such

¹ NCLB also requires performance breakouts by gender, for low-income students and for students with learning disabilities.

factors, and their individual contribution to the gap, is unclear. If, for example, the district leadership knows that teacher interactions, early reading skills, and parental involvement are all potential drivers of the achievement gap, and they have limited time and resources to investigate solutions, how should they prioritize their efforts? Although the drivers may remain the same across geographies, their relative influence may not.

1.2. OBJECTIVE

In this dissertation, we employ a variety of statistical methods that provide a bridge between large-scale studies of achievement gaps and the analyses necessary to address the needs of a single community. We believe that community-based, small-scale studies are necessary because they allow us to connect general trends to local behavior. Although the achievement gap is an issue of national concern, local action is required to affect change, and, although there are many factors that may predict achievement gaps, the prevalence of each factor may vary substantially from one district to the next.

For the local school leaders who face these issues, it is necessary to prioritize strategies for reform, and it is our belief that quantitative studies using local data contribute analytical tools and insight that will help guide these decisions. Contemporary studies of achievement gaps have generated a large set of hypotheses for explaining the achievement gap; the purpose of this work is to assist school districts in testing various hypotheses through rigorous analysis of local data. The goals of our efforts here are to:

- 1) Improve local understanding of the dynamics and magnitude of local achievement gaps; and to
- 2) Inform the district of the potential effectiveness of a given strategy, assisting their efforts to reduce achievement gaps.

This thesis focuses on analyzing the data about achievement gaps for a particular school district. Per this limitation, there are a number of potentially relevant factors that we do not

address, such as parental involvement and peer pressure. Our work focuses on the quantitative aspects of achievement gaps, and some of the relevant factors are inherently less quantifiable, or even unquantifiable. Given these limitations, we recognize that data analysis is not the sole means to understanding achievement gaps; however, we believe that data analysis is a critical component in developing coherent, objective strategies for reducing the gaps.

The school district that participated in our research possesses a number of appealing qualities for this work. In many ways, the district and its surrounding community are atypical, meaning that it is the kind of community traditionally overlooked in studies regarding achievement gaps. The district's students, as a whole, perform well above state averages on standardized tests. The district has a substantial minority population, which aids our ability to draw insights regarding minority performance. Also, the district is located in a middle-class suburb, an environment in which factors such as poverty and crime are less prevalent. Most important, the district maintains an especially rich set of longitudinal achievement data, which affords us the opportunity to consider multiple years of data for a given group of students. With few exceptions (e.g. Ferguson, 2001; Ogbu, 2003), achievement gaps within middle class, highly integrated communities are a relatively unexplored area of research.

1.3. FINDINGS

Our guiding motivation was to determine “what the data had to say” regarding the magnitude, growth, and associated trends concerning the local achievement gap. The analyses reflect performance in Mathematics; a subject for which we have the most extensive dataset and a subject which been cited by some authorities as a “clearer measure of school effectiveness”² than others that are available. Our primary data source was standardized test data, collected over a period of 7 years (1999 – 2006); aside from the

² Schemo, 2006.

state exam, there was also access to middle school classroom data, such as math grades and teacher assignments. As the study progressed, the data revealed several insights involving differences in achievement among two groups of district students: a *minority* group, which is mostly comprised of Black students, and a *majority* group, largely comprised of White students. In summary:

- **The study confirmed for the district that there has been a consistent and significant difference in the Math performance of majority and minority students.** Conventional means of measuring achievement provide the first indication of a performance gap in the district. However, no single measure of achievement is infallible, as even the most common approaches to measuring achievement appear prone to mischaracterizing achievement gaps under certain circumstances. In response to these limitations, we developed a series of alternative metrics which compare group performance in a *relative* and an *absolute* sense; relative measures focus on the question of *who's ahead*, and absolute measures focus on *how far* one group is ahead of another. Regardless of what metric we used, however, we found an appreciable gap in Mathematics achievement between majority and minority students. This gap was apparent at all grades between the second and the eighth, and within three different cohorts of students. Relative gaps were generally larger in magnitude than absolute gaps,
- **Apparent constancy in the overall achievement gap with age can conceal growing gaps among students of similar skill.** In an analysis of student data spanning Grades 2 through 8, our relative and absolute metrics reveal only slight changes in the magnitude of the overall achievement gap as the students moved through elementary school. However, when we compared trends in progress among students with similar third-grade performance, we found that minorities at all skill levels were consistently less likely to improve their level of performance with respect to the testing standards.

- **External factors believed to predict minority achievement are present, yet their influence on the gap is unclear.** Minorities in the district have a higher mobility rate than the majority group; however, there is little difference in the eighth-grade performance of minorities who have been in the district for several years versus recent transfers. In a similar vein, minority households have a lower income distribution than majority households; however, achievement gaps exist throughout the income distribution, and there is little difference in the performance of low-income minorities and their more well-off colleagues.
- **In this district, boys make less progress than girls over time, regardless of race.** An analysis of performance with respect to gender indicates that minority males have comparable performance to minority females in Grade 3, and they lose ground to minority females in Grades 3 to 8. In contrast, majority males initially outperform females in Grade 3, before the females catch up in Grade 8. Although minority males have the most room for improvement, initiatives that address the needs of both genders will be necessary in closing the eighth-grade gap between races.
- **Early interventions might provide the best opportunity for the district to reduce achievement gaps.** Shortly before this study began, the district had launched a program aimed at improving reading comprehension in the earlier grades. Under the assumption that improved reading proficiency could lead to larger math gains, eliminating the reading gap before sixth grade could reduce the eighth grade Math gap by 24%. In sixth grade, students in the district transfer to middle school, where high-performing students place into advanced math classes. Over 70% of the majority group places into the advanced math courses, as opposed to less than 30% of the minority group. By sixth grade, most minority students do not have the grades to qualify for advanced placement, and among those who do,

minority students are less likely to enroll. Although there is evidence to suggest that the district could narrow the 8th grade gap by increasing Honors enrollment among qualified minority students, the low numbers of qualified minorities mitigate the estimated opportunity.

- **Testing data may provide insights into teacher effectiveness.** Using student “peer groups”, defined by performance, ethnicity, and math course, we developed a rating system to assess the relationship between middle school teachers and the relative gains made by their students. Most teachers received roughly the same ratings and had similar observed success in both ethnic groups. However, there were a handful of teachers associated with exceptionally high (and low) student gains. Also, when we compared teacher effectiveness with respect to race, most teachers rated comparably with both ethnic groups, providing little evidence of teacher bias based on ethnicity. In short, the best teachers appear to benefit students of all races: there is little evidence of a teaching style that benefits one race far more than the other.

1.4. OVERVIEW OF CHAPTERS

Following this introductory chapter, **Chapter 2** provides a brief review of the Achievement Gap. There is a vast literature relating to achievement gaps; our review will focus on the aspects most relevant to our work. As we will show, there are a number of ways to measure achievement gaps; in **Chapter 3**, we address the question of characterization. This chapter provides a review of the most prevalent approaches to measuring gaps, and it also outlines a number of alternative metrics designed to provide a more nuanced view of student achievement.

The next two chapters concern the magnitude and dynamics of the achievement gap in the district studied. After an overview of the school district, **Chapter 4** examines the eighth grade achievement gap during the 2004-05 academic year. We focus on those students who

have been in the district for at least five years or more; this group referred to as the *2005 cohort*. The discussion involves testing gaps on the state exam, as well as grade gaps in the classroom. From there, **Chapter 5** explores the dynamics of the district gap over time. We compare the 2005 cohort data with results from the previous and subsequent cohorts, and we look backward in time to observe how the gap has changed as students have progressed through the district.

With an understanding of how the magnitude of the achievement gap has changed, the next chapters look to district data to determine how some specific patterns within the district might predict the trend. In **Chapter 6**, the focus is on elements of the student's external environment; specifically, we consider district data regarding student mobility and economic data. We also focus on the performance of minority males in the district and compare their progress to the outcomes for minority females, and majority students. In **Chapter 7**, the focus shifts to the student's school experience; here, we use data to evaluate various strategies for narrowing the gap. We assess the correlation between reading comprehension and math gains, and we also consider the role of middle school honors-course placements in predicting achievement on the 8th grade exam. Finally, we make some observations regarding teacher effectiveness with respect to Math gains. The final chapters offer some commentary regarding the findings. In **Chapter 8**, we recount some important caveats to consider regarding the work presented here and in **Chapter 9**, we summarize our work and offer some thoughts on future directions for this research.

2. Background

This chapter provides a brief introduction to the existing literature about those aspects of the achievement gap most relevant to this thesis. Within the broad literature regarding race and achievement, our study addresses a particular, yet critical question: What can a district learn about achievement gaps from its own data?

2.1. Historical Perspectives on Race and Achievement

2.1.1. The Coleman Report (1966)

Of the many studies and articles written regarding race and academic achievement in the United States, a report titled *The Equality of Educational Opportunity*, published in 1966, remains among the most influential. Commissioned as part of The Civil Rights Act of 1964, the report was national in scope, with a sample size of approximately 600,000 students in over 4,000 schools. The study, also known as “The Coleman Report” after sociologist and lead investigator James S. Coleman, established methods for measuring student achievement that persist to this day. In a reflection on the influence of the report, Coleman expressed his belief that the study’s scope and visibility helped establish the practice of “evaluating schools in terms of their results [i.e., student achievement], rather than their inputs.” (Coleman, 1972)

The study presented three major findings (Kahlenberg, 2001):

- 1) Funding disparities between “black schools” (i.e., schools composed mainly of black students) and “white schools” were relatively small, given the achievement gap;

- 2) The economic status of a student's family was more predictive of student achievement than school funding; and
- 3) The economic status of a student's classmates appeared to have an effect on student achievement, above and beyond the student's personal economic status.

Prior to the study, many researchers (including Coleman) believed that the achievement gap was largely the result of differences in the funding of "black schools" and "white schools". However, the findings of that study refuted that hypothesis and motivated further study of how factors other than funding predict achievement and achievement gaps. As such, the findings of the Coleman Report serve as a basis for most of the research that has followed.

2.1.2. The Persistence of Achievement Gaps

Since the release of the Coleman Report, national data suggest that, despite some early improvement, race-based achievement gaps have not gone away. Since 1970, the Department of Education has administered the National Assessment of Education Progress (NAEP) as a means of tracking national trends in educational achievement. The NAEP is a series of standardized tests that serve as the national barometer for academic achievement.

We illustrate the general trend in NAEP data with an example. Since the initial Mathematics assessment, given in 1973, data from the NAEP Long-Term Trend exam indicate that the difference in average score between 13-year old blacks and whites fell consistently until the mid-1980s, when progress in closing the gap stalled (Figure).

Test Year	Average NAEP Score		
	White	Black	Gap
1973	274	228	46
1978	272	230	42
1982	274	240	34
1986	274	249	25
1990	276	249	27
1992	279	250	29
1996	281	252	29
1999	283	251	32
2004	288	262	26

Figure 1: NAEP Long-Term Trend (LTT) Mathematics Exam Data, 1973 – 2004

As the figure shows, regular increases in the average performance of black students explained much of the initial improvement. As noted by Kober (2001), the NAEP data show similar patterns over this period among students in different age groups, and on exams in different subjects (e.g., Reading comprehension). Thus, despite some progress, achievement gaps have remained a critical issue at the national level.

With the passage of The No Child Left Behind Act of 2001 (“NCLB”), closing achievement gaps became an explicit goal of US education policy³. The stated objective of NCLB is to ensure that all US schoolchildren meet or exceed the academic standards set forth by the states in which they reside. To monitor states’ progress toward this objective, NCLB requires regular testing and score reports for several student subgroups, defined by factors such as race, gender, financial status, and English proficiency.

By holding schools accountable for subpar performance within subgroups, NCLB makes explicit the need to schools to improve the outcomes of all students, regardless of race. In a letter dated April 23, 2007, U.S. Secretary of Education Margaret Spellings released a policy letter regarding the reauthorization of NCLB. After meeting with a number of chief

³ In addition to the achievement gap, the NCLB Act calls for the removal of achievement gaps across the economically disadvantaged and students with learning disabilities.

state education officers, she reaffirmed that the first priority of her reauthorization proposal would be to “strengthen efforts to close the achievement gap” in elementary and secondary education.⁴

2.2. KEY ISSUES REGARDING ACHIEVEMENT GAPS

2.2.1. External Effects versus School Effects

One of the main findings of The Coleman Report was that factors external to the school system, such as social class and parental education, predict much more of the variation in student performance than certain school-based factors, such as school funding. As noted, the Report’s findings inspired a large body of subsequent research on school effects. Although there has been some criticism of the methods used in the Coleman study, the majority of studies in this arena⁵ have supported the notion that, on a national level, achievement gaps are reflective of socioeconomic gaps between Whites and non-Whites (Murnane et al., 2006).

Perhaps the strongest argument for the dominant influence of external factors comes from the wealth of evidence that race-based achievement gaps appear in the earliest years of schooling. Historically, Black people have been at a disadvantage to Whites on a host of socioeconomic indicators. With regard to racial minorities in general, it is common to refer to people categorized as Black, Hispanic, or Native American as “underrepresented minorities”, to denote their lack of proportional representation in the middle and upper-classes of American society. The “underrepresented minority” designation generally excludes people of Asian descent, a minority group that has reached a certain level of socioeconomic parity with Whites.

⁴ Letter retrieved from <http://www.ed.gov/policy/elsec/guid/secletter/070423.html> on May 29, 2007.

⁵ Murnane et al refer the interested reader to Hanushek (2003) for an overview of the evidence.

As it turns out, this social clustering with regard to race is consistent with the race-based achievement gaps. For example, on the 4th Grade NAEP Mathematics General exam⁶ (Figure 2), we see that White and Asian students have consistently outperformed Blacks, Hispanics and Native American students. Further, recent studies of achievement gap among younger students (e.g., Bali and Alvarez, 2004; Fryer and Levitt, 2005) indicate that, for black and white students, achievement gaps appear as early as Kindergarten, which is the start of formal education for many children. Although educators are accountable for closing the achievement gaps, it is apparent that gaps exist when children first enter the classroom.

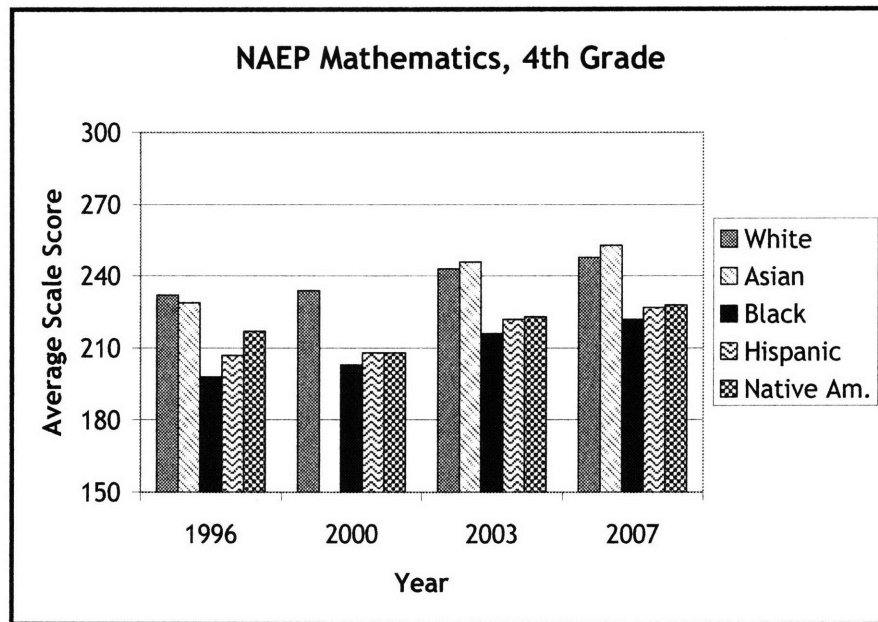


Figure 2: Average NAEP Score, Grade 4 Mathematics (Source: Nation’s Report Card, 2007⁷)

Given the strong influence of environment, the ability of schools to close the achievement gap has been, and continues to be, a controversial topic. As students progress through school, evidence at the national level suggests that the achievement gaps among ethnic

⁶ The data in Figure 2 are from the NAEP General Exam, as opposed to the NAEP Long-Term Trend Exam (c.f. 2.1.2).

⁷ 2000 Data for Asian students omitted from NAEP reports due to concerns regarding data accuracy and precision.

groups are present from the beginning of schooling, and, ultimately, grow larger over the course of elementary and secondary school (Phillips, Crouse, and Ralph, 1998). According to Cain and Watts (1970), the objective of The Coleman Report was to describe various aspects of the US educational system and analyze how they relate to achievement, with the objective of prescribing policies to improve the system. However, the Report's conclusions led many theorists, including fellow sociologist Seymour Martin Lipset to conclude that, "schools make no difference; families make the difference."⁸

2.2.1.1. School Quality and Achievement Gaps

Over time, reflection on the original Coleman findings and later studies have provided more nuance to the study of school effects, ultimately reviving the notion that schools do matter. In a recent study, Hanushek and Rivkin (2006) provide a broad analysis of the contemporary black-white achievement gap. The study addresses two recurrent questions regarding achievement, such as:

- 1) How large of a role does school quality play in the size and growth of the (black-white) achievement gap?
- 2) Which factors of school quality correlate highest with student achievement?

In what follows, we review the methodology and findings on a few specific topics, namely:

- Changes in achievement gap from Grades 3 through 8
- Achievement gap trends with respect to prior performance
- Key correlates of achievement

These are just a few of the topics Hanushek and Rivkin address in the study; however, we call attention to these topics because of their direct relevance to our study and methods of inquiry.

⁸ As recounted in (Hodgson, 1973). Lipset was a prominent sociologist in his own right; along with Coleman and Martin Trow, he co-authored the book *Union Democracy*.

The researchers primarily base their analyses on data from the Texas School Project (TSP), a repository of school characteristics and student performance data for students attending public elementary schools throughout the state⁹. The TSP sample consists of over 400,000 students, with black students representing 26% of the sample. Each student in the sample is linked to a series of records that extend from Grades 3 to 8. The dataset consists of four consecutive cohorts of students; the earliest cohort completed 8th Grade in 1999, and the most recent cohort completed 8th Grade in 2003.

The measure of performance is the Mathematics portion of the Texas Assessment of Academic Skills (TAAS). The TAAS is a criterion – based test, with scoring scales that vary across grade levels. To account for these variations, the researchers use standardized TAAS data, resulting in achievement measures with zero-mean and unit-variance at each grade level.

Using TSP data, the authors present the black – white achievement gap at a given grade level as the difference in the (standardized) mean score for blacks and whites. Using this approach, the researchers found that the black-white achievement gap grows over time. In 3rd Grade, the mean TAAS score for blacks was 0.59 standard deviations lower than the mean for white students; in 8th grade, the gap had grown to 0.70 standard deviations. As we explain further in **Chapter 3**, the standardization of achievement data can distort the magnitude of group differences in achievement in several ways. The authors do not claim otherwise, nor do we dispute their general findings; however, we note that standard deviation is not a fixed unit of measure, and is very easily misinterpreted.

Early in their study, Hanushek and Rivkin indicate their preference of studying cohorts of students, rather than the full TSP dataset. At the time of the study, the full TSP dataset

⁹ In addition to TSP data, the researchers also included analyses from the Early Childhood Longitudinal Study (ECLS). The ECLS data, which is a national sample of data collected between Kindergarten and 5th Grade, has no direct relationship to the TSP data; in this study, it is used to assess early achievement gaps, particularly in response to Fryer and Levitt (2005).

contained approximately 500,000 whites and 140,000 blacks at any given grade level; however, the cohort sample consisted of about 337,000 whites and 88,000 black students. Although the cohort approach reduces the total number of records and in this case, the proportion of black students, use of cohorts ensures that the measured changes in the gap are occurring among a specific set of students, a more reliable basis for measurement.

Achievement gap trends with respect to prior performance

Building upon their overall assessment of the gap, the researchers next consider the growth in the achievement gap among students with comparable 3rd grade Mathematics skills. A stated reason for doing so was to test the premise that achievement gaps “may grow more rapidly for initially high-achieving blacks.” The approach calls for dividing the data into subgroups (representing different levels of 3rd grade skill), measuring the black-white achievement gaps within each subgroup, and comparing the results.

Notably, the researchers argue *against* the idea of using the 3rd grade Math scores to define the subgroups for 3rd grade Math skills. Instead, the researchers use 3rd grade *Reading* scores as a proxy for initial Math skills. Their methodology is based on the assumption that Reading scores and Math skills are positively correlated, and that using Math scores would increase the likelihood of errors in measuring the “true” change in the achievement gap.¹⁰ Comparing trends within subgroups, the researchers found evidence that growth in the achievement gap correlates with initial skill. In the subgroup with the lowest 3rd Reading scores, the gap went from grew from 0.51 to 0.58, increase of 14%. However, in the subgroup with the highest 3rd grade Reading scores, the gap grew from 0.18 to 0.44, an increase of 175%.

¹⁰ The increased likelihood of error stems from the fact that a given Math score is actually an *estimate* of true Math proficiency, subject to measurement error, and the best (i.e., error-minimizing) estimate of true proficiency, for a member of a group, is the group mean. Since the mean score for blacks is, in general, lower than the mean score for whites, blacks with high 3rd grade scores are more likely to have their scores inflated by measurement error, and are therefore, more likely to under-perform against a high scoring white 3rd grader on a subsequent test. By a similar argument, low scoring whites would be more likely to have scores deflated by measurement error, and would be more likely to outperform a black student with the same score on a later exam. Either scenario would result in an overstatement of the growth in the achievement gap, due to initial classification errors.

Despite their reservations, the actual risk of using initial Math scores as a proxy for Math skills might be overstated. Although the researchers assume a correlation between 3rd grade and Reading scores and 3rd grade Math skills, they provide no guidance regarding the requisite degree of correlation. In the event of weak correlation between Reading scores and Math skills, the substitution might well introduce classification errors greater than the ones it attempts to prevent.

Key Correlates of Achievement

The researchers propose a regression model for relating student achievement to the various data tracked by the TSP. The response variable, denoted A_{iG} , represents the achievement test score (A) for a given student (i) in a given Grade (G), attending school s_{iG} . They model A_{iG} as the result of a broad array of factors, grouped under the following categories¹¹:

α_{iG} : Intercept Term

F_{iG} : Family and Student Factors (e.g., race, income)

$P_{iG s_G}$: Peer Factors (racial composition, % low-income)

$T_{iG s_{iG}}$: Teacher Factors (Education, tenure)

$S_{iG s_{iG}}$: Non-teacher student factors (class size, expenditures)

e_{iG} : Error term

Due to the size of the TSP dataset, the authors chose not to perform the regression using student level data. Instead, the researchers aggregated student data, creating meta-records on the basis of race, grade level, academic year, and school attended.¹² An implication of the data aggregation is that the focus of the regression shifts to comparisons across classrooms rather than across students. Even then, the regression is based on a blended composite of all classrooms at a given level. The regression will therefore highlight differences in “good classrooms” and “bad classrooms”, losing any of the variations that might occur within a given grade level.

¹¹ Although the authors do not provide a full list of the factors used in the regression analysis, we highlight the structure of the regression to frame the discussion.

¹² To illustrate, one meta-record might contain all information pertaining to *black students* at “School X” attending 8th grade in 1999.

Key Correlates of Mathematics Achievement, Texas Public Schools*				
Race School Level	Black		White	
	elementary	middle	elementary	middle
% new to school	x	x	x	x
% black	x	x		x
% first-year teachers	x	x	x	x
% second-year teachers		x	x	

* Adapted from Hanushek, Rivkin (2006) fixed effects regression model
x : statistically significant ($t > 1.96$)

Figure 3: Key Correlates of Student Achievement, Texas Public Schools

From the regression, the researchers identify four factors with statistically significant effects on achievement (Figure 3). The proportion of students new to the school is a peer-based factor, reflective of student mobility within a given school. The proportion of black students is also a peer factor, indicative of the racial composition of the school. The final two components relate teacher experience to student achievement; schools with high proportions of teachers with little or no prior experience tend to produce lower test scores.

In summary, the study finds that black students tend to go to schools with higher rates of mobility, more inexperienced teachers, and higher populations of black students, and the researchers find these characteristics most responsible for the growth of the achievement gap. However, due to highly aggregated data, the applicability of these results to a given school district is unclear.

2.2.2. Achievement Gaps among the Middle-Class

This thesis is an examination of achievement and race in a racially integrated, middle-class suburb, an environment with a fairly specific geography and demographic. Within such a specific demographic, many of the socioeconomic factors that are present in the aggregate

may not differ greatly for blacks and whites. In that sense, studies of this nature offer perspective on whether closing socioeconomic gaps would be sufficient to close the achievement gap.

The late anthropologist John Ogbu was among the first to examine achievement gaps in middle class communities. In his study of the Shaker Heights (Ohio) school system, Ogbu (2003) reports significant differences between black and white student performance. Given the strong academic reputation of the district, the study underscored the importance of disaggregating achievement trends within school districts that are generally viewed as “high-performing.”

2.2.2.1. GPA Gaps in Shaker Heights (Ferguson, 2001)

Although there have been relatively few studies of achievement gaps among the middle – class, Ferguson (2001) provides a quantitative analysis of black-white achievement trends in Shaker Heights, a suburb of Cleveland, Ohio. Much like the setting for our study, Shaker Heights is a highly diverse, middle class community. In what follows, we provide a brief review of the study, focusing on the data, methods, and findings as they relate to our study.

Whereas the Hanushek/Rivkin study focuses on differences in school quality as a predictor of widespread achievement gaps, Ferguson looks to differences in student culture for insight regarding the local achievement gap. Given a number of characteristics that represent the various backgrounds, beliefs, and attitudes of the student body, the study address the following questions:

- 1) Which characteristics vary significantly with respect to race?
- 2) How well do these characteristics predict achievement?

Characteristics that vary significantly with race *and* predict achievement are the ones most critical to closing achievement gaps.

Data

The data for this study come from student responses to a survey, The Cornell Assessment of Secondary School Culture. The district administered the survey in the spring of 1999 to students ranging from grades 7 to 11. Although the study contains responses from across the district, there are no observed controls for grade level. Of the 1,699 students completing the survey, 1382 (81%) identified as either white or black. Indicative of the racial makeup of Shaker Heights, the district has nearly equal proportions of black and white students; the data contain 685 black students and 697 white students¹³.

The survey is designed to characterize influential elements of student culture by collecting information across a variety of different categories. Categories of the survey include:

- Race and Family Background (e.g., no. of parents/siblings in the household)
- Reasons for Not Studying/Completing Homework (e.g., other commitments)
- Classroom Attitudes and Behaviors (paying attention in class, helping others)
- Other Attitudes and Behaviors (e.g., # of hours watching television)
- Perceptions of Popularity (e.g., which characteristics define popularity)

The primary measure of achievement is the student's most recent GPA¹⁴. Across all grade levels, students report black-white differences in GPA of approximately one letter grade. Among males, blacks had a GPA of 2.1 versus 3.2 for whites. Among females, black students had a GPA of 2.4 versus 3.4 for whites. Students also reported the proportion of advanced level courses they were taking, the amount of time spent on homework, and their homework completion rate. In a sense, these data represent secondary measures of achievement.

¹³ These figures are reported in Table A-1 of the Ferguson study.

¹⁴ Students receive grades of A-F; student GPA is based on a 4-point scale with A=4 points, B=3 points, and so on.

Notably, the study considers a number of factors that school districts do not regularly track. This consideration reflects the focus of this study, which relates to differences in student culture, rather than differences in school resources. Conversely, many of the data maintained by the district, such as teacher credentials or standardized test data are not considered.

Methods

The paper uses regression analysis throughout to identify relationships in the data. The regression models follow two general forms:

- 1) Student attitudes and behaviors, as predicted by race, gender, and family background.
- 2) Achievement metrics, as predicted by race, gender, family background, attitudes, and behaviors

Models of the first sort serve to identify statistically significant differences in cultural norms across blacks and whites. Models of the second sort serve to identify the aspects of culture most indicative of student achievement.

The regression models for predicting GPA incorporate over thirty predictor variables to account for survey responses, grouped according to the categories of the survey. In order to perform the regression, Ferguson uses a combination of approaches to quantify the survey data. A few data, such as GPA and number of hours devoted to homework, enter the regression model without modification. Several data items come scaled response, or Likert, items¹⁵; these data (or composites thereof) appear in standardized form, a process that transforms raw data to a dataset with zero-mean and unit standard deviation. The regression model uses binary variables to identify classification data such as gender, race,

¹⁵ In this context, a Likert item consists of a statement (e.g., “I pay attention in school”), and a numerical scale. Survey respondents use the scale to rate the applicability of the statement. For example, the preceding statement might be associated with a scale ranging from 1 (“totally disagree”) to 5 (“totally agree”).

household composition and parental education; survey questions with “yes/no” or “agree/disagree” answers also require binary variables.

When constructing regression models from survey data, a certain level of data manipulation (e.g., use of binary variables, or standardization) is usually required. As regression models become more extensive, the level of manipulation increases, and with each manipulation, we run the risk of distorting relationships within the data.

In this context, the conversion of scaled response data to standardized equivalents carries a particular risk of distortion. For ease of exposition, assume that all scaled response items are to be rated on a scale of 1 to 5. When tabulating the data, it is conceivable that student response would vary considerably on some items, and hardly at all on other items. The recognition that some items are more polarizing than others might carry significance of its own, yet when we apply a common distribution (zero mean and unit deviation) to each response item, the differences in variance from one item to another is lost.

Findings

Consistent with the study’s focus, Ferguson provides a number of findings regarding race, culture, and achievement. Many of the findings, though provocative, lie outside of the scope of our study. Therefore, we do not review all of the findings here. However, we do highlight a few results regarding achievement.

By multiplying the regression coefficient for a characteristic by the mean black-white difference, Ferguson estimates the extent to which said item predicts achievement. According to the data (Figure 4), a student’s percentage of advanced courses is the single best predictor of GPA, predicting about 25% of the GPA gap among males (0.28/1.13), and 28% of the GPA gap among women (0.26/0.93).

Key Predictors of GPA in Shaker Heights, Ohio*		
Factor	Male	Female
Parents education	0.11	0.11
Household composition	0.05	0.05
Attitudes and Behaviors	0.11	0.08
Pct. Of Advanced coursework	0.28	0.26
Homework Completion Rate	0.11	0.11
Homework Hrs. /night	0.00	0.01
Total Predicted Difference	0.66	0.61
Total Actual Difference	1.13	0.93
Predicted /Actual	59%	66%

* Adapted from Ferguson (2001) extended regression model

Figure 4: Key Predictors of Student GPA (Ferguson, 2001)

After advanced coursework, parental education, student attitudes and behaviors (in and out of the classroom), and homework completion rate were the next best predictors of GPA. Differences in household composition (i.e., the number of parents/siblings in the home) are, by comparison, a relatively weak predictor of GPA. Across all indicators, a substantial amount of the GPA gap remains unaccounted for by the regression model.

2.2.2.2. Additional findings from Middle School

After Shaker Heights, Ferguson analyzed results from a broader survey, conducted during the 2000-2001 school year by the Minority Student Achievement Network (MSAN). The MSAN study surveyed over 40,000 students attending school in Shaker Heights, and over a dozen other middle to upper class school districts throughout the US.

In his analysis of the MSAN data, Ferguson (2002) again noted evidence of significant differences in achievement across ethnic groups in middle-class communities. This time, Ferguson incorporated three different measures of achievement: student GPA, student

comprehension of course material, and comprehension of textbook and other study material.

In his study, Ogbu attributed the difference in outcomes to cultural differences regarding the value of education, and more specifically, to a lack of parental guidance and student interest in high performance (Lee, 2002b). Although Ferguson noted persistent gaps in middle-class communities, he did not find evidence to support Ogbu's hypothesis regarding a diminished interest in educational success among minorities. To the contrary, Ferguson found that attitudes toward education, and time spent on homework, were similar across ethnic groups. Despite the latter point, minority students had a lower homework completion rate, lending credence to the possibility of a "skills gap" among minority students.

Also, there was evidence to suggest that underrepresented minorities respond to certain teacher behaviors differently than their peers in the majority. For example, Black and Hispanic students appear more motivated by teachers who "encourage" rather than "demand" academic performance. In contrast, the White and Asian students appeared to be equally motivated in either case.

2.2.3. Measurement and Reporting

When studying achievement gaps, the metric used to measure achievement can vary greatly. When available, there is also the option of using classroom data, like GPA, or student-reported measures, such as student comprehension. Alternatively, one could interpret achievement as the attainment of some academic standard, such as high school completion and college enrollment.

Most often, education researchers look to standardized test scores for measures of academic performance. Standardized tests come in a variety of forms, and the method of reporting performance can vary as well. Boudett, City, and Murnane (2006) identify three general types of standardized tests, each with their native reporting methods:

Norm-referenced tests (NRTs) are concerned with gauging a student's performance relative to the performance of a larger, peer group. A student's percentile rank is a common method of reporting performance on an NRT. The SAT college exam is a well-known example of an NRT

Criterion-referenced tests (CRTs) are concerned with testing a student's mastery of a specific body of knowledge. A CRT will typically have a cutoff score which represents a basic level of competency, and students either pass or fail the exam. In such an example, the CRT measure would be the percentage of students who passed the exam.

Standards-referenced tests (SRTs) are concerned with a student's performance with respect to a set of knowledge standards. SRTs are a more complex form of a CRT, as the standards represent differential levels of mastery. For example, a math student could be judged to have Basic, Proficient, or Advanced mastery of the material.

In practice, there is considerable overlap in the design and reporting of standardized tests. In accordance with the NCLB Act, all states currently use some form of SRT to measure K-12 student achievement. Using these tests, the U.S. Department of Education requires states to report a standards-referenced metric; namely, the percentage of students meeting or exceeding academic standards. However, on a student's individual performance report, norm-referenced measures (such as percentile rank) are also given.

When analyzing achievement data, the choice of reporting metric has been shown to influence one's perception of group performance. Seltzer, Frank, and Bryk (1994) analyzed achievement on a reading exam using two different metrics; one norm-referenced and the other criterion-referenced. The study was an analysis of reading scores for a cohort of students, taken from Grades 1 through 6. The objective was to determine whether the

change in achievement was greater from Grade 1 through 3, or Grades 4 through 6. The norm-based measure indicated faster growth in the early years, whereas the criterion-based measure indicated that growth accelerated in the later years. Although both metrics were derived from the same data, the indicators yielded conflicting results regarding rates of progress.

2.3 OUR CONTRIBUTIONS

The studies mentioned above represent contributions to a collective understanding of how and why achievement gaps can occur. Student achievement, and hence the achievement gap, is an educational outcome that is perhaps best thought of the combination of several factors, both internal and external to the school system.

While the precise contribution of school factors is a matter of debate, studies like (Hanushek and Rivkin, 2001) submit evidence that suggests that schools can play a significant role in closing achievement gaps. However, as we have noted, trends reflected in national achievement data do not necessarily correlate to the needs of any given community, a discrepancy that can hamper the ability of local school leaders to focus on the needs of their students.

Studies of middle-class communities offer the opportunity to examine achievement gaps within a particular community. Thus far, such analyses have offered perspectives on how racial and gender differences relate (or fail to relate) to observed achievement gaps. However, the data used to capture these local insights have no direct connection to the metrics used to evaluate the district. Hence, there is little guidance to local educators how to use their own data to understand local achievement.

Regarding methodology, we note that most studies heavily favor the use of regression analyses, data standardization, or both. Like all methods, these approaches make certain assumptions about data, which have the potential to distort the analysis if the assumptions

don't hold. For this reason, educators might benefit from the application of alternative, complementary methods of data analysis.

To that end, this thesis presents a number of analyses designed to help school leaders develop a richer understanding of student achievement data. The thesis also develops a series of tools for testing various hypotheses about local achievement gaps. In total, we view the following contributions as a complement to the large and ongoing body of research devoted to achievement gaps; in particular, our work provides an analytical toolset for local educators who wish to learn from the achievement trends of their students.

Our contributions address the measurement, characterization, and investigation of local achievement gaps. Regarding metrics, we present multiple metrics for measuring relative and absolute achievement gaps in group performance. The presented metrics are applicable to samples of arbitrary size and distribution, which makes them especially suitable for studying achievement trends at the local level. Next, we develop a detailed characterization of achievement trends for the district, utilizing a series of analyses to test for evidence of several hypotheses regarding the Oak Park achievement gap.

Finally, we investigate the prevalence of several factors that may correlate with high achievement. We develop several models that reflect different reform scenarios that the district might attempt to implement. Also, we develop a novel approach to consider student gains as a proxy of teacher effectiveness. Regarding the last contribution, and all of the analyses herein, we accept that test data and course grades connect to a great many factors predicting achievement, and that our findings are not definitive. However, we do expect that the analyses help us determine what the data have to say about achievement in Oak Park.

3. Testing and Measurement

The term “achievement gap” implies a significant difference in academic performance between two groups. However, the perception of an achievement gap depends on how one measures it. This chapter begins with an overview of three of the most prevalent approaches to measuring achievement gaps. As we will see, there is no one “right” way to measure achievement gaps, as each approach has its limitations to consider.

The limitations found in existing metrics motivate the proposal of several alternative measures of the gap in the later half of the chapter. These alternative metrics complement, rather than unseat, existing measures of the gap; our intent is to develop a comprehensive understanding of group outcomes, drawn from multiple interpretations of data. Among the proposed alternatives, we distinguish between absolute and relative measures of achievement gap, noting that both types of comparisons play a role in understanding differences in achievement.

3.1. MEASURING ACHIEVEMENT GAPS

We begin with a look at how achievement data are currently used to measure achievement gaps. Conceptually, an achievement gap is the difference in the academic performance of two (or more) groups. The size of the achievement gap depends on a number of things, including but not necessarily limited to:

- How we measure individual performance;
- How we aggregate individual data to define group performance; and
- How we choose to compare the performance of the groups.

Each of these factors can influence our perception of an achievement gap, and as we review the following approaches, we will see how these factors can aid or hinder our understanding of the gap.

3.1.1. Average Score

Given a list of student test data, a direct approach to expressing group performance would be to take the group average. Commonly, the group average is reported as the arithmetic mean; however, there are other interpretations, such as the median, which may also be used to represent average performance. Regardless of how we choose to define the group average, the achievement gap, in this approach, equals the difference in average score.

For example, suppose we wanted to compare the exam performance of two groups. Group A has an average exam score of 90, and Group B has an average score of 80. If we wanted to present the achievement gap as a comparison of averages, we would report a 10-point gap between Groups A and B, with Group B trailing Group A. Based on this information, how much do we actually know about the magnitude of the achievement gap?

A limitation of measuring and reporting the gap in terms of scoring scales is that one requires additional knowledge about the exam to make sense of the information. In the example above, it is impossible to grasp the severity of the achievement gap based on the given information. As noted by Boudett et al. (2005), testing scales are generally arbitrary in structure, and as a result, the meaning and interpretation of a test score can vary greatly from one exam to the next. As a result, it is nearly impossible to make sense of a measure based on scale scores without some degree of familiarity with the test.

3.1.2. Performance Levels

Testing scales can be somewhat complicated, so it usually helpful to have alternative representations of student performance in order to provide some context. For state-

administered exams, the prevalent alternative to reporting scale scores is to describe achievement in terms of performance levels.

Per the requirements of the federal No Child Left Behind Act (NCLB), state exams are designed to measure student proficiency in a specific set of academic content and skill areas, which we will refer to as learning standards. To make the connection between scores and standards explicit, the scoring scale maps to a series of performance levels, (There are usually 3 – 4 performance levels for a given test.) The performance levels describe exam performance with respect to the learning standards; for a group of students, this mapping allows us to categorize students by performance level, and use the levels as our measure of group performance

For example, the National Assessment of Educational Progress (NAEP) maps exam scores to one of three performance levels; Basic, Proficient, and Advanced. Conceptual definitions of the NAEP performance levels are given in Figure ; as an alternative to using average NAEP scores to describe group achievement, a commonly used approach is to report the percentage of students at or above the “Proficient” performance level.

Basic	Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
Proficient	Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Advanced	Superior performance.

Figure 1: NAEP Performance Level Definitions¹⁶

Performance levels link explicitly to the scoring scale with *cutscores*. The cutscore is the lowest allowable exam score within a given performance level; for example, on the NAEP exam, there is a cutscore which separates *Basic* performance from *Proficient* performance,

¹⁶ Accessed from <http://nces.ed.gov/nationsreportcard/mathematics/achieve.asp>

and another cutscore which separates *Proficient* performance from *Advanced* performance. The placement of the cutscores depends on the grade level and subject area; in Figure 2, the cutscores are labeled *C1* and *C2*.

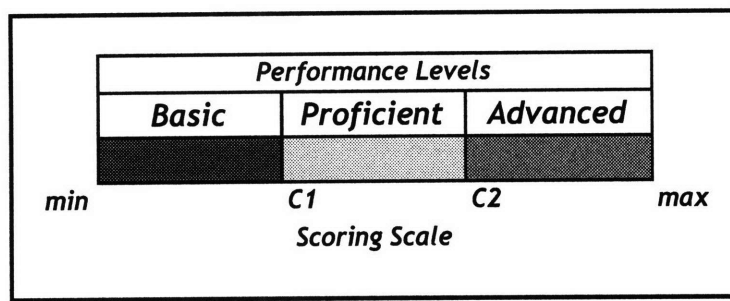


Figure 2: Depiction of Performance Levels and Cutscores

Performance levels allow us to describe student achievement without the added complexity of testing scales, and by doing so, they simplify the reporting of group achievement. By replacing score data with performance levels, the findings are freed from the context of a single exam, and are made accessible to people unfamiliar with the nuances of the given test. Also, the concept of subject matter proficiency provides a common framework for comparing test data from different exams (e.g., proficiency in mathematics vs. proficiency in writing skills).

Given the role of performance level metrics in reporting group achievement, the metrics also facilitate the reporting of achievement gaps. Suppose that the “percentage of students meeting or exceeding standards” is the preferred way of measuring group achievement. Given two groups (say, A and B), we can state the achievement gap as the difference in the percentage across the groups. For example, with the knowledge that 90% of the students in Group A met or exceeded standards, and 80% of Group B met or exceed standards, we can report a 10% achievement gap between the two groups.

Limitations of Performance-Level Metrics

Metrics based on performance levels simplify the task of expressing achievement gaps, but unfortunately, these metrics can also conceal important details about group performance. Performance levels condense a range of student performance into a handful of categories, a process that discards potentially valuable data.

The potential for misinterpreting data is especially high near the cutscores. Suppose we have a student (“Student X”) whose exam score is just above the cutscore for *Proficient* performance and another student (“Student Y”) whose exam score just below the cutscore for *Advanced* performance. As shown in Figure 3(a), although both students performed at a *Proficient* level, the scale scores clearly indicate that Student Y outperformed Student X. Although it may be fair to say that the students attained the same level of performance, it would be misleading to say that there was no achievement gap amongst the students. Thus, we see that performance-level metrics can miss achievement gaps that occur within performance levels.

Conversely, performance level metrics might also overstate achievement gaps. In Figure 3(b), we depict a scenario in which Student X attains a score slightly lower than the *Basic/Proficient* cutscore, and Student Y attains a scale score slightly higher than the cutscore. In terms of performance levels, Student Y is placed among the *Proficient* students, and Student X is not.

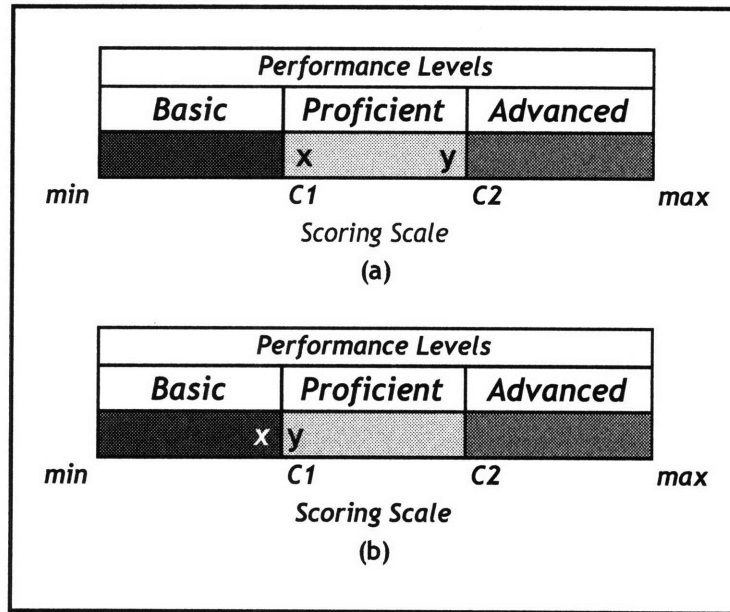


Figure 3: Divergent Representations of Achievement Gaps

This simplified view of relative achievement tells us that Student Y outperformed Student X, but it does not capture the fact that the difference in scale scores between X and Y is relatively small. Further, had Student X scored lower on the exam (e.g., near the minimum), the difference in performance levels would remain unchanged; this is in spite of the fact that the achievement gap, in terms of scale points, would be larger. In this instance, the performance-level metric would be too crude a measure to accurately depict achievement gaps.

To be clear, our intent is not to diminish the importance of performance levels. Although the metric has some shortcomings as a measure of achievement gaps, the performance levels do provide a quick, intuitive method for summarizing group performance. Also, the performance benchmarks serve as a state-sanctioned mechanism for grouping students into different categories of academic skill; a property that will be of use throughout this study.

3.1.3. Standardized scores

Scale scores and performance level are, arguably, the foremost indicators of an individual's performance. However, as we have shown, both metrics have limitations in the context of studying achievement gaps. Scale scores, in general, only apply to the exam of record, and can be easily misinterpreted by those unfamiliar with the structure of the exam scale. Performance level metrics do not require knowledge of scoring scales, but they summarize group data in a way that can mischaracterize achievement gaps.

In response to these limitations, many studies of the achievement gap rely on a modified dataset as the basis for their analyses and findings. A popular modification, known as *standardization*, applies a linear transformation to exam scores, creating a dataset with zero mean and unit variance. When this transformation is applied to normally (Gaussian) distributed data, the resulting dataset has what is known as the *standard* normal distribution, hence the name.

For a collection of test scores with mean μ and standard deviation σ , the necessary transformation is below. If a student achieved a score of X , that score maps to a standardized equivalent, denoted $Z(X)$ via the following transformation:

$$Z(X) = \frac{X - \mu}{\sigma}$$

We will refer to the standardized value $Z(X)$ as the **Z-score**. By construction, the *Z-score* reinterprets student performance as a measure of how well the student performed relative to the group average. With a standardized dataset, we measure performance in standard deviations; that is, a Z-score of 1.0 would indicate that a student's test score was one standard deviation higher than the group average.

Given a set of student test scores, we can standardize test performance for any given student using the approach described above. Furthermore, standardization is also to compare achievement across groups. In addition to the overall mean and variance, suppose

that we also knew that the average test scores for two mutually exclusive subgroups (A and B) within the data. Just as before, we could express the performance of Groups A and B in terms of deviations from the collective mean (μ):

$$Z(\text{Group A}) = \frac{\mu_A - \mu}{\sigma}; \quad Z(\text{Group B}) = \frac{\mu_B - \mu}{\sigma}$$

In the equation, μ_A and μ_B represent the mean scores from Groups A and B, respectively, and σ represents, as before, the standard deviation of the full population.

Now, with standardized representations of the performance of both groups, the difference in performance is simply the difference in the Z-score across groups, which simplifies to the following expression:

$$\text{Standardized Difference in Means} = \frac{\mu_A - \mu_B}{\sigma}$$

The metric is a representation of the achievement gap between Groups A and B, measured in standard deviations. In the event that the two groups are normally distributed and collectively exhaustive (that is, if all test takers in the dataset either belong to Group A or Group B), this difference in means metric is equivalent to the Student's t statistical test for independent means. Under those same conditions, the metric is also equivalent to Cohen's d , which measures effect size across groups (Cohen, 1988)¹⁷.

3.1.3.1. Benefits of Standardization

In some respects, the use of standard deviation as a measure of achievement is an improvement over the previously discussed methods. As we have already seen, test scores can be difficult to interpret without specific knowledge of the scoring range. Once a set of student scores have been standardized, technical knowledge of the underlying scoring scale, such as the range and calibration, is no longer necessary. Instead of relying on some arbitrary scale, the standardization approach instructs the reader to think of performance

¹⁷ Achievement gap studies that concern two disjoint groups generally meet this condition. For example, in a study of the achievement gap between male and female students, any difference in normalized group performance would be interpreted as the empirical "effect size" of gender on student performance.

gaps in terms of “standard deviations”, a statistical concept which is fundamental to data analysis.

In general, the standard deviation is a measure of data dispersion; however, the standard deviation takes on added significance for Gaussian data. When data are normally distributed, the standard deviation links to stronger statements about the distribution. For example, regardless of the nominal values of the mean and standard deviation, when data are normally distributed, approximately 95% of observations will fall within 2 standard deviations of the mean, and 99% of observations will fall within 3 standard deviations.

Under the assumption that the data are normally distributed, standardization also facilitates the development of complementary measures of performance. For example, a student’s *percentile* rank (i.e., the percentage of students in the sample with a lower score) relates directly to the *Z*-score. For example, consider the percentile rank of a student with a *Z*-score of 1.0. This number represents the percentage of students with *Z*-scores of less than one. Alternatively, the percentile represents the likelihood that a student, chosen at random, would have a *Z*-score less than 1.0. If the data are normal, this likelihood is equivalent to the normal CDF (cumulative distribution function) of 1, $\Phi(1)$.¹⁸

3.1.3.2. *Limitations of Standardization*

Gaussian Assumption

Although many have adopted the practice of measuring achievement gaps in terms of standard deviations, the method is not without its limitations. As we have noted, *when data are normally distributed*, the standard deviation allows us to make broader statement about the full distribution. However, in empirical work, the assumption of normality may not necessarily apply.

¹⁸ Tables which approximate the cumulative distribution function of a normal random variable, $\Phi(z)$, are readily available.

To illustrate, suppose there is an achievement test given over a large population of students, scattered across several school districts, and we have the task of analyzing score data for a single district. Although it may be convenient to assume that the district data will be normal, there is no compelling reason to make this assumption *a priori*.

Even when the score distribution of the full test-taking population is Gaussian by design, we can not assume that district data will be normal; for the district of interest need not be statistically representative of the full population. In a study of a larger scope, perhaps with a larger number of districts and students, it is conceivable that the distribution of test scores would begin to resemble a normal curve, but in an analysis with a non-representative subset of the test-taking population, it is important that the metrics used are suitable for the data.

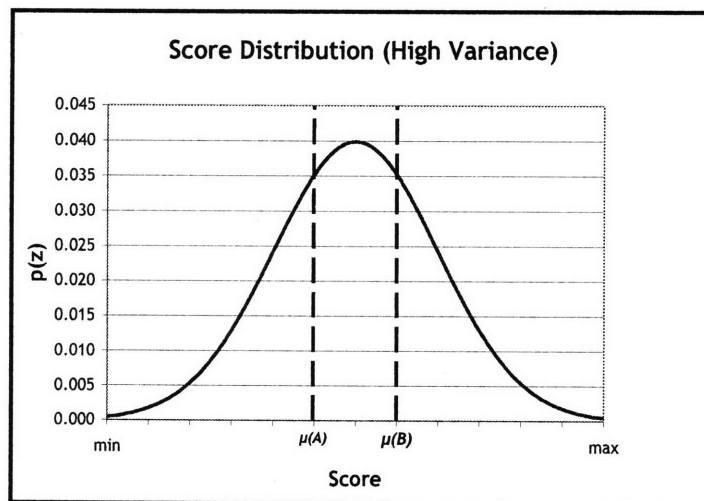
Ambiguity of Standard Deviation

Assumptions of normality aside, the standard deviation is a somewhat ambiguous unit of measure. The standard deviation, in itself, does not represent a fixed measure of difference; rather, its magnitude directly follows from the group variance. Thus the significance of a gap measured in standard deviations is subject to change with the group variance. In a dataset with low variance, an achievement gap of several deviations may translate to a rather modest nominal gap.

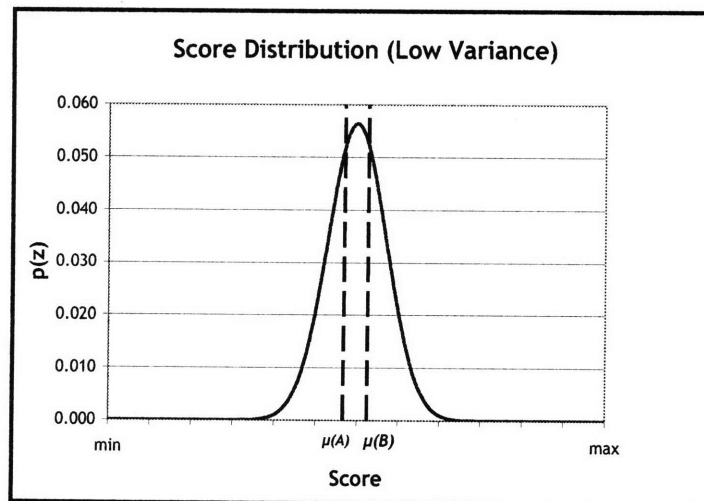
Suppose we are interested in comparing the academic performance of two groups of students (Group A and Group B). If the Group A mean is one standard deviation lower than the Group B mean, what does this say about the severity of the achievement gap between the two groups? The charts in Figure 4 depict two scenarios in which we depict a combined distribution of scores from Group A and Group B over a fixed range¹⁹.

¹⁹ For ease of exposition, assume that, in both scenarios, the combined test scores are normally distributed, the combined mean is in the middle of the scoring range, and that Groups A and B are of equal size.

Now, in both cases, the Group A mean is one standard deviation less than the Group B mean; however, it is somewhat misleading to suggest that the achievement gap between the two groups is identical. Arguably, one might conclude that the gap is smaller in the second scenario, as the means for Group A and Group B are much closer. Without a broader understanding about the magnitude of the standard deviation, we are left with a unit of measure that, in the words of Barnett (1995), might function “less like a trusty yardstick than a distorting mirror.”



(a)



(b)

Figure 4: One Deviation, Two Distributions

Sensitivity to Group Size

In comparing the achievement gap of two groups, it is unclear that the relative size of the groups should matter. However, when we combine groups, the relative proportion of the groups will certainly affect the variance of the combined group. As a result, a measurement based on standard deviations will also be sensitive to the relative size of the groups of interest.

We can illustrate this sensitivity, and its effect on measuring achievement gap in standard deviations, with a simple example. Suppose there is a population composed of two groups of students. On some achievement test, we'll assume that all of the students in Group A receive a score of 80, and that all of the students in Group B receive a score of 100. As a result, there is a fixed difference in means of 20 scale points.

Now, to find the standardized difference in means, we divide the difference in means by the combined variance, σ^2 . To find variance, we use this formula:

$$\sigma^2 = E[X^2] - (E[X])^2$$

In the formula, X represents a scale score, chosen at random from the combined population. The first term is the expected value of X^2 , and the second term represents the expected value of X , squared. As we will show, we can raise or lower the variance simply by changing the ratio of Group A and Group B students, thereby altering the standardized gap metric.

When Groups A and B are of equal size (Figure 5), the probability of choosing a student from either group is 0.5. Therefore, a score chosen at random has an equal probability of being 80 or 100. In this case, the expected value of X is $(0.5 \times 80) + (0.5 \times 100) = 90$. Also, the expected value of X^2 is $(0.5 \times 80^2) + (0.5 \times 100^2) = 8200$. Using the above formula, the variance is $(8200 - 90^2) = 100$ and the standard deviation (σ) is 10. As a result, the gap between Group A and B equals 2 standard deviations (20/10).

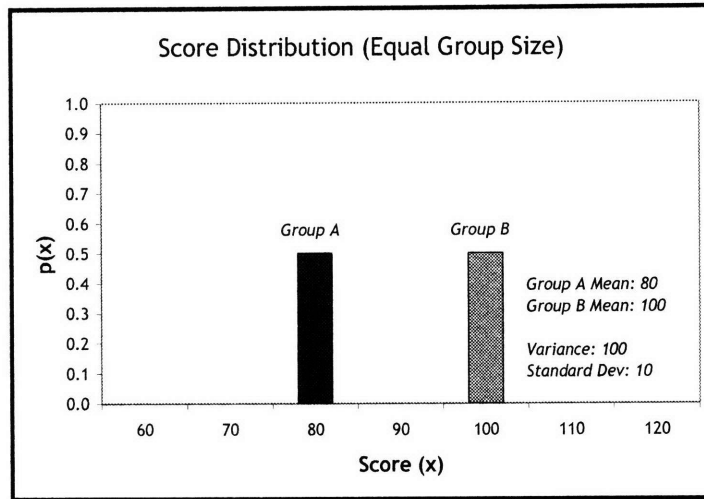


Figure 5: Variance among Equal Group Sizes

Next, we assume that groups are of unequal size. In Figure 5, only 10% of the students belong to Group A, and the remaining 90% belong to Group B. In this case, the expected value of X is $(0.1 \times 80) + (0.9 \times 100) = 98$, and the expected value of X^2 is $(0.1 \times 80^2) + (0.9 \times 100^2) = 9640$, resulting in a variance of $(9640 - 98^2) = 36$ and a standard deviation (σ) is 6. Without changing the nominal difference in scores, the standardized gap between Group A and B increases from 2 to 3.33 standard deviations ($20/6$).

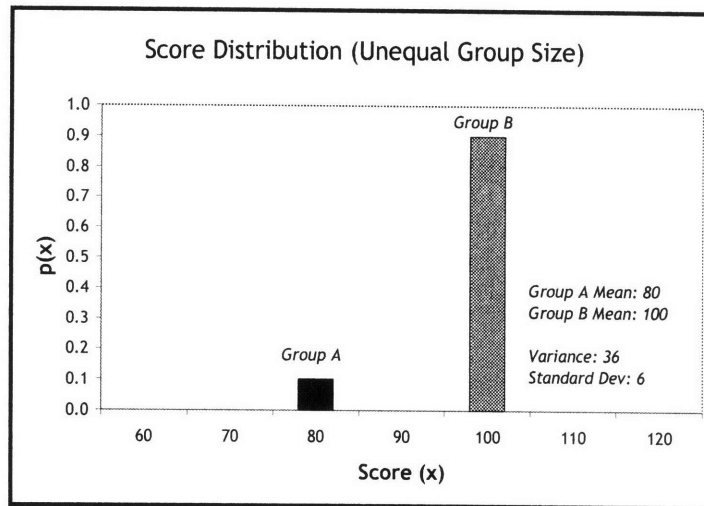


Figure 6: Variance among Unequal Group Sizes

Although there is considerable support for the practice measuring achievement gaps in terms of standard deviations, we have noted circumstances in which the standardized metric can distort our interpretation of the achievement gap. The magnitude of the standard deviation varies with the distribution of the student body; as a result, the metric can lead to ambiguous conclusions about the true nature of the gap. As we have said before, our intent is not to discredit the metric, but to raise awareness concerning its limitations.

3.2. ALTERNATIVE MEASURES OF ACHIEVEMENT GAPS

In the previous section, we note that there are multiple ways to measure the achievement gap. Among the more prevalent approaches, we find that each method has its benefits and limitations. With that point in mind, we are motivated to consider a number of alternative methods for measuring achievement gaps.

In keeping with the subject of this study, each of the metrics discussed here are applicable to populations of arbitrary size and distribution. Among these alternatives, we distinguish measures that compare actual performance outcomes (*absolute* measures) from measures

that effectively rank the performances of the two groups (*relative* measures)²⁰. Relative measures help us gauge which group is ahead, whereas absolute measures help us gauge how far one group is ahead of the other.

For consistency, all of the metrics introduced in this section are indexed on the interval $[-1, +1]$, a scale which aligns with the familiar statistical concept of correlation. An index value of -1 denotes a scenario least favorable to one of the groups, and a value of $+1$ denotes a scenario least favorable to the other group. In all cases, an index value of zero indicates no observed net difference in performance of the groups.

3.2.1. Relative Achievement Gaps

We will begin with a discussion of three relative measures of achievement gap. We refer to these metrics as relative metrics because they are concerned specifically with the relative position of members from each group. We construct these metrics from differences in the student rank and distribution, rather than nominal differences in test scores. Because of their construction, the following metrics rely strictly on local data – information about the scoring ranges, or the calibration of the scale, is not required.

3.2.1.1. *The Median-as-Percentile Index (MPI)*

The Median-as-Percentile Index measures achievement gaps by expressing the median performance of one group of students as a percentile of another group's performance. Similar metrics have prior usage in illustrating group differences in achievement, behavior, and resources; for example, Herrnstein and Murray (1994) use this approach to describe differences in the learning environments of Blacks and Whites in *The Bell Curve*.

²⁰ To be precise, *absolute* measures are a subset of *relative* measures in the sense that their computation requires more information about the data; namely, the actual scores and a relevant range of potential scores. Here, we use the terms to signal the difference between rank-based and nominal measures of performance gaps.

In general, the metrics we use measure differences in two populations, or groups; for ease of exposition, we'll refer to these groups as the *primary* and *secondary* groups, respectively. Given two groups, the designation of a primary and secondary group is arbitrary. However, when there is a hypothesis that one group will be at a disadvantage, it is customary to treat the (supposedly) disadvantaged group as the primary, and use the median of that group as the basis for comparison. For example, in the context of racial gaps amongst black and white students, one might say:

“The median test score for black students corresponds to the n^{th} percentile of test scores for white students.”

Using the median for the primary group as a benchmark, the corresponding *percentile* for the other group is a number between 0 and 100; low percentiles indicate that the primary group is lagging behind the secondary group, whereas high percentiles would indicate that the primary group is outperforming the secondary group. A percentile of 50 would indicate that the two groups had the same median, and presumably no performance gap.

To create the *MPI* metric, we represent the percentile value on an interval of [-1, 1] via linear transformation:

$$MPI = \frac{\text{Percentile} - 50}{50}$$

By construction, percentile values lower than 50 result in negative values of *MPI* for the primary group.

We illustrate the calculation of *MPI* with an example. Suppose we have test data for two groups of students: Group A and Group B, with Group A as the *primary* group. There are 10 students in each group, and we rank the test scores in each group from lowest to highest, as in Figure 7. Group A is our primary group, and the table shows that the median of the Group A data $((48 + 49)/2 = 49.5)$ is higher than only one of the Group B scores.

Student	Group A	Group B
1	13	46
2	14	51
3	16	53
4	27	57
5	48	60
6	49	66
7	62	69
8	83	72
9	94	75
10	95	100

Figure 7: Ranked Test Scores for Two Groups

The magnitude of *MPI* captures this relationship between the Group A median and the scores in Group B. The median is larger than 1 of the 10 scores in Group B, placing the median in the 10th percentile of Group B scores. Using the formula, we would calculate an *MPI* of -0.80 $\{= (10 - 50) / 50\}$ on a scale of -1 to 1.

The *MPI* metric can highlight stark differences in the distribution of outcomes of two groups, but we note that the *MPI* metric measures the achievement gap by comparing two distributions at a single point (i.e., the median of the primary group). By design, the *MPI* metric locates the “average” (i.e., median) performance of one group within the performance distribution of a comparison group. Although the median is arguably a critical point in the distribution, a metric drawn from a single point may leave us with an incomplete view of the two distributions.

One drawback of the metric is that the magnitude of *MPI* can change depending on the choice of the primary group. For example, suppose we use the data in Figure 7 to calculate *MPI* again, this time with Group B as the primary. The Group B median is 60, which corresponds to the 60th percentile for Group A. Using the formula, the *MPI* is $(60-50)/50 = 0.20$. When we compare this to our earlier result (-0.80, with Group A as the primary), the

magnitude of the gap, from the perspective of Group B, is much smaller, leading to a conflicting view of the severity of the gap.

3.2.1.2. *The Rank-Sum Index (RSI)*

The Rank-Sum Index shares its name with the Wilcoxon rank-sum statistical test²¹. The purpose of the test is to determine whether two sets of data share a common distribution. Conceptually, the test measures the overlap in two distributions by ranking the scores in the combined distribution, and then evaluating the sum of ranks, or *ranksum*, for one of the groups (i.e., the *primary* group); in what follows, we adapt the test statistic to measure achievement gaps.²² As in our discussion of *MPI*, we will assume that we are comparing the performance of Groups A and B, and that Group A is the primary group. However, unlike the *MPI* metric, the choice of primary group does not affect the magnitude of *RSI*.

Given test score data from two groups, with Group A as our primary, we calculate the *ranksum* as follows. We rank scores from the full dataset in increasing order; thus, the lowest score in a distribution of N scores has a rank of 1, the second-lowest score has a rank of 2, and so on up to N ²³. The ranksum is literally the sum of the ranks belonging to the primary group. By design, the ranksum is largest if every student in Group A outperforms, or outranks, every student in Group B. Conversely, the ranksum is smallest if every student in Group B outranks every student in Group A.

An Example of the Rank-Sum Index

We introduce the rank-sum index with an example, depicted in Error! Reference source not found.. Suppose we have test scores of 10 students. The students belong to two groups: Group A has 3 members, and Group B has 7 members. In order to compare the results of the two groups, we would first rank the students from 1 to 10, based on their test score.

²¹ The Wilcoxon rank-sum test is equivalent to the Mann-Whitney U test.

²² Lieberman (1976) proposed a similar method of adapting the rank-sum test purpose of measurement.

²³ In the presence of ties, we assign average ranks. For example, if the two lowest scores were tied, then both scores would receive a rank of 1.5 (i.e., the average of ranks 1 and 2)

The *RSI*, as mentioned earlier, is a measure of the overlap in the distribution of two groups. In doing so, the *RSI* also describes a probabilistic result. Assuming that we have Groups A and B, with Group A as the primary group, we use the *RSI* to approximate the likelihood that a student from Group A outperforms a student from Group B. Using linear interpolation, we convert *RSI* to a new quantity, ρ , defined on [0,1]:

$$\rho = \frac{RSI + 1}{2}$$

Now, let a^* and b^* represent randomly chosen scores from Groups A and B, respectively. It can be shown that the quantity ρ is also equal to the following expression:

$$\rho = P(a^* > b^*) + 0.5 \times P(a^* = b^*)$$

In the absence of ties in the data, this expression equals the probability that a random score from Group A is higher than a random score from Group B, which is equivalent to the likelihood that a student in Group A outperforms a student in Group B. When we introduced the *RSI* metric, we presented a scenario that yielded an *RSI* of -0.52 ; using the equation above, we could also say that the likelihood of a student from Group A outperforming a student from Group B was 0.24

3.2.1.3. The Kolmogorov – Smirnov Index (KSI)

Like the Rank-Sum Index, the Kolmogorov – Smirnov Index takes its name from a statistical test for comparing two populations of data. Given data from two populations, the two-sample Kolmogorov – Smirnov (KS) test also evaluates the null hypothesis (i.e., assumption) that the datasets have the same underlying distribution. Where the rank-sum test measures the overlap between two distributions, the Kolmogorov-Smirnov test focuses on differences in cumulative distribution. The test statistic, put simply, is the maximum “distance” between the distributions of two sets of data.

To determine the KS test statistic, we compare the empirical cumulative distribution function (CDF) of both groups. For a given set of student test scores R , the CDF of a particular scale score i is defined as the probability that a student picked at random would have a test score less than or equal to i :

$$F(i) = P(r \leq i | r \in R)$$

For the student with a scale score of i , the CDF is nearly equivalent in meaning to the student's *percentile* within that group. Given a range of possible scores R , we define F_A and F_B as the CDF for Group A and Group B, respectively. The KS test statistic (K), then, is equal to the maximum vertical difference between the plots:

$$K = \max_{i \in R} \{F_A(i) - F_B(i)\}$$

To provide some intuition for K , notice that, for every value in R , the metric compares the fraction of Group A test scores that are i or lower with the same fraction for Group B. If the two distributions are essentially the same, then none of the comparisons should stray too far from 0, and thus neither should K .

With the KS test statistic in hand, only slight modification is need to construct the metric KSI for measuring achievement gaps. Like all of the gap metrics in this section, KSI will take values on the interval $[-1, 1]$. Without loss of generality, we will let positive values represent comparisons that favor Group A, and negative values represent comparisons that favor Group B.

Now, the magnitude of the test statistic lies in the interval $[0, 1]$, and is equal to the magnitude of KSI . To determine the sign of KSI , let i^* represent a value for which the difference in distributions is maximized; that is, $|F_A(i^*) - F_B(i^*)| = K$. Typically, the

sign of KSI is determined by the position of the CDF plots at i^* ; if F , the CDF for Group A, is larger at that point, then KSI is negative, otherwise, KSI is positive.

Assuming that positive values of KSI favor Group A, we compute KSI from the KS test statistic as follows:

$$KSI = \begin{cases} K & \text{if } F_B(i^*) > F_A(i^*) \\ -K & \text{if } F_B(i^*) < F_A(i^*) \end{cases}$$

To determine KSI , it is helpful to plot the CDFs of the groups in question. In our discussion of the MPI metric, we incorporate example data from Groups A and B; in Figure 9, we have plotted the CDF data for the two groups. The scoring range forms the x-axis, and the y-axis is the CDF. For example, if we want to determine the likelihood of getting a score larger than 80 on the exam, the plot indicates 70% of students in Group A had a score of 80 or lower, as compared to 90% of students in Group B.

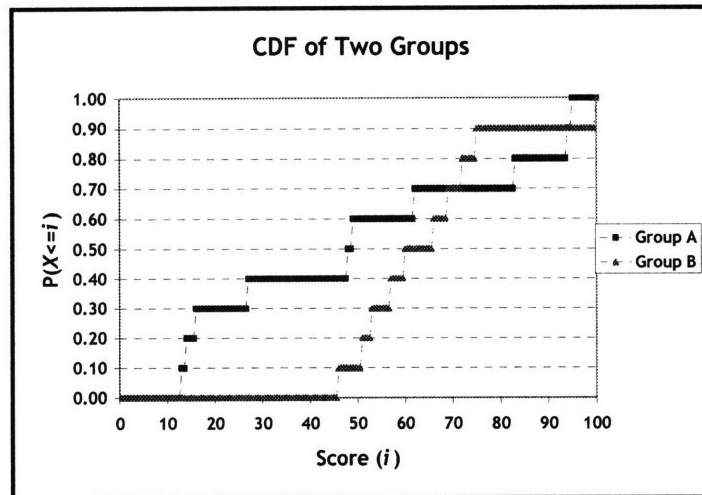


Figure 9: CDF Plots of Two Groups

Now, the difference in CDF is largest for a score of 50; 60% of Group A has a score at or below 50, as opposed to only 10% of Group B students. The KS test statistic, then, is the absolute difference in CDF, which is $|0.60 - 0.10|$, or 0.50. Noting that the CDF is smaller

for Group A (i.e., our primary group), the *KSI* for our example is -0.50, indicating that Group A is at a disadvantage.

3.2.2. Absolute Achievement Gaps

Aside from considering the relative differences in group performance, it is also helpful to understand the sheer size of an achievement gap. Even if one group completely outperforms the other, the difference is not consequential if it represents a one-point gap on a scale of 100. To suggest the magnitude of the overall gap between groups, we rely on absolute gap metrics.

Although relative and absolute measures are related, differences in the development of absolute metrics can lead to divergent perspectives of the severity of the achievement gap. As in the case of the relative metrics, we propose three absolute metrics, all defined on a [-1, 1] scale. We calculated all of the absolute measures from actual scale score data and the associated scoring range. The scoring range is important because it represents the range of attainable scores for all test takes. By dividing the difference in scale score by the range of all scores, we calculate the magnitude of the gap as a percentage of the largest attainable gap.

Regarding these metrics, we make no assumption about the shape of the distribution; instead, we choose multiple perspectives of “average” group performance to mitigate the likelihood of misinterpreting the extent of the absolute gap. If the metrics converge, then we have an indication that the group distributions are similar. Conversely, divergent metrics would indicate that the distributions have different shapes

3.2.2.1. Absolute Mean Gap

Perhaps the most common approach for summarizing group performance is to find the group mean – i.e., calculate the sum of all data points, and divide by the size of the group.

Assuming that we have Groups A and B, with Group A as the primary group, the *mean gap* is calculated as the difference in means (Group A – Group B), divided by the length of the attainable scoring range. We let X represent the attainable scoring range; the term $\max(X)$ denotes the highest possible score, and $\min(X)$ denotes the lowest.

$$MeanGap = \frac{\mu(A) - \mu(B)}{\max(X) - \min(X)}$$

By construction, the above transformation converts the difference in means to a value on the $[-1, 1]$ interval. We used the same convention for our relative metrics, but, we interpret the magnitude of absolute gaps in a different way. If the value of an absolute gap metric reaches -1, this would imply that every student in the primary group scored the minimum, and that every student in the other group scored the maximum.

3.2.2.2. Median Gap

As an alternative to the mean, the group *median* is another interpretation of average group performance. Although the mean is most commonly associated with the notion of “average” performance, the median is not as sensitive to outliers; a property which, in some cases, might make the median a more “stable” metric for consideration. For instance, if a single student in a group scores well above (or below) the distribution of the group, that student’s performance has no impact on the group median.

We represent the median using the inverse CDF function, F^{-1} . Recall that the CDF calculates the percentage of students at or above a given score level; conversely, the term $F^{-1}(0.50)$ denotes the median – that is, the score in the distribution for which the likelihood that a student is below that score equals 50 percent.

$$MedianGap = \frac{F_A^{-1}(0.50) - F_B^{-1}(0.50)}{\max(X) - \min(X)}$$

When the test scores in each group are normally distributed, the difference in the arithmetic mean and median (50th percentile) gaps is negligible. However, as mentioned earlier, there is no reason to assume *a priori* that the achievement data are normally distributed.

3.2.2.3. *Quartile Gaps*

The mean and median present two approaches to describing the middle of the data; however, the statistics say little about what is happening within the two halves of the distribution on either side of it. In a further attempt to compare differences in the distribution of minority and majority achievement, we incorporate these halves into a third measure of absolute gaps – the quartile gap.

The quartile gap is determined by comparing data quartiles; numbers which represent the 25th and 75th percentiles of a distribution. We compute the quartile gap by finding the average of the 1st and 3rd quartiles for both groups, and then dividing the differences of the averages by the length of the scoring range.

$$QuartileGap = \frac{0.5 \times (F_A^{-1}(0.75) + F_A^{-1}(0.25)) - 0.5 \times (F_B^{-1}(0.75) + F_B^{-1}(0.25))}{\max(X) - \min(X)}$$

The quartile gap metrics considers differences in the distribution that occur away from the center of the distribution. By averaging the first and third quartile, we create another perspective of average group performance. If the group distributions have a similar shape, then the quartile gap will be roughly equal to the median gap; however substantial differences in the size of the quartile gap would indicate divergent trends in the tails of the distribution.

Regarding Absolute Measures

Collectively, these measures provide different perspectives of the “average” difference in performance. The three absolute measures clearly have a similar structure, so why bother

with the additional measures? In our desire to develop a more nuanced view of student performance, we allow for the possibility that the subtle variations in these measures of the gap may reveal divergent trends in student performance.

Although each of the metrics compares average performance of the two groups, the metrics have their strengths and weakness. The mean is arguably the most intuitive metric, yet outliers heavily influence the mean of a distribution. The median gap is resistant to outliers, but group medians do not provide much insight into the shape of the distribution. When we add the quartile gap, we create a more comprehensive picture of the differences in distribution; together, the three metrics allow us to create a composite view of the scoring gap at various points across the distribution.

3.3. ABSOLUTE METRICS VS. RELATIVE METRICS

The difference in an absolute measure and a relative measure of the performance gap lies in the context. Relative gap measures are concerned exclusively with how two groups compare to one another. With regard to the population under study, relative gap metrics define the gap in terms specific to the local data; hence, the focus is on rank order and student percentiles. In contrast, absolute measures look at the local achievement gap in more general terms, use globally defined standards of performance, such as the scale score or a performance benchmark.

To understand the effect of these divergent perspectives on measuring the gap, it is helpful to consider scenarios which occur at the extremes of our $[-1, 1]$ scale. An extreme relative gap corresponds to a scenario in which one group dominates the other group (e.g., every student in Group X has outperformed every student in Group Y). In contrast, a extreme absolute gap does not only indicate that one group has dominated another, but that the two groups occupy opposite ends of the score distribution (i.e., Group X outperformed *all* other groups, and Group Y was bested by *all* other groups).

Although both scenarios are highly unlikely, understanding the difference is essential to recognizing that our perception of achievement gaps, and public perception of achievement gaps, is dependent on how we measure them. For example, suppose that a school district releases a report that the difference in average test score among boys and girls was thirty points on a scale of 200. It is conceivable that the difference in score would provoke some concern in the community. However, if the district also reported that the lowest scoring member of one group outperformed the highest scoring member of the other group, this additional information might provoke a stronger response.

3.4. SUMMARY

There are a number of ways to measure achievement gaps. In the preceding chapter, we discussed some of the more prevalent methods, and noted that every approach has its benefits and limitations. With that in mind, we proposed six alternative approaches to complement existing methods the achievement gaps. All the alternative methods were indexed on a $[-1, 1]$ scale, with a score of 0 indicating no evidence of an achievement gap.

In our discussion of achievement gaps, we differentiate between relative and absolute measures of achievement gaps. Relative measures provide a perspective of *who's ahead*, based solely on the rank order of test scores, whereas absolute measures reflect *how far apart* the groups are, in terms of the structure of the exam. In our view, relative and absolute measures complement each other, and we believe that both types of measures inform the study of achievement gaps. In the next chapter, we employ these metrics as we begin to explore the achievement data of a middle-class elementary school district.

4. The 8th Grade Mathematics Gap in Oak Park

In this chapter, we use the metrics discussed in **Chapter 3** to analyze achievement data from an actual school district. Using data about student ethnicity, we compare the academic outcomes of two groups of students; a majority group, comprised mostly of White students, and a minority group, largely comprised of Black students. Before doing so, we provide a brief overview of the school district and our primary measure of student achievement, the Illinois State Achievement Test, or ISAT. We also introduce a method for recalibrating the ISAT scoring scale as a means of expressing standards performance with a continuous metric.

We begin our analysis with data collected from eighth graders during the 2004-2005 academic year. The metric of choice for our analysis concerns Mathematics, a subject area which has been cited as a stronger indicator of school effectiveness than, say, reading comprehension²⁴. In keeping with our interest in the local district, we focus on the outcomes of students who have spent at least five years in the district, a group we refer to as “veterans” of the district. Our insights regarding eighth grade Math performance consider two perspectives: test performance on the ISAT; and classroom performance, in the form of course grades.

²⁴ Schemo, 2006.

4.1 DISTRICT OVERVIEW

The data referenced throughout this study belongs to Oak Park Elementary School District #97, also known as “OP97”. The district serves the village of Oak Park, Illinois, which is a suburb on the western border of Chicago. In economic terms, Oak Park is a middle-class community; according to Census data collected in 1999, Oak Park residents had a median family income of \$81,703, with a poverty rate of 5.6%. As of late 2007, Oak Park has a population of approximately 52,000 residents, with a district enrollment of approximately 5,000 students²⁵.

For the purposes of this study, we make a distinction between the performance of White and Asian students – hereafter referred to as the *majority (MAJ) group* – and the performance of Black, Hispanic, and Native American students, collectively referred to as the *minority (MIN) group*.²⁶ As the table below indicates (Figure), over 90% of Oak Park students fit into either the MIN or MAJ group category. Regarding the remainder, we can not reliably assign multiracial students to either group; thus, we limit our analyses to a comparison of the MIN and MAJ groups.

²⁵ The population figure comes from the official Oak Park website, and the district enrollment figure comes from district literature. Estimates were obtained on October 16, 2007.

²⁶ Due to the nature of the study, multiracial students are excluded from the study.

OP97 Student Body Composition (2006)		
Ethnic Group	Race	Pct. Of Total
MAJ Group	White	55.9
	Asian/Pac. Islander	3.5
		59.4
MIN Group	Black	29.2
	Hispanic	4.0
	Native American	0.1
		33.3
	Multiracial	7.4

Figure 1: OP97 Student Body by Ethnicity (2006)

Although the MIN and MAJ groups, as defined, contain students of multiple ethnicities, Figure makes it clear that the MAJ group is mostly comprised of White students, and the MIN group is mainly composed of Black students. Therefore, although we combine students from different ethnic backgrounds, the differences observed between majority and minority students, are essentially the differences in the performance of White and Black students.

4.2 THE ISAT MATHEMATICS EXAMINATION

Every spring, students in Oak Park are required to take the Illinois State Achievement Test (ISAT). The ISAT, which was first introduced in 1999, is used by the state Board of Education to assess student knowledge with respect to state-approved education goals, known collectively as the Illinois Learning Standards (ISBE, 2003). By extension, test results are used by the state to assess the effectiveness of schools and school districts at teaching these standards. In accordance with the “No Child Left Behind” Act of 2001, federal assessments of statewide educational performance are also based on the outcomes of this test.

This study makes use of ISAT data collected from 1999 through 2006, inclusive. However, prior to the 2006 examination, a number of critical changes were made to the structure, administration, and scoring system of the ISAT Math exam. Thus, to accommodate these changes, test scores recorded before 2006 were converted to the current scale using tables made available by the state board (ISBE Bridge Study, 2006).

4.2.1. Scoring the ISAT exam

The students in this study were given the ISAT Math examination three times, in Grades 3, 5, and 8. The content of the exam varies by grade, that is, the 8th Grade exam is designed to test knowledge of concepts taught during 8th grade, and is therefore, a more difficult exam than, say, the 3rd grade exam. However, a single vertical scale is used to calibrate the scores for all three exams. As shown in Figure 2, the minimum score on the Grade 3 exam is 120, and the range of possible scores increases with each grade level. The use of a single scale for all three exams implies that an exam score of, say, 240 is meant to represent a specific level of performance, regardless of whether the score was attained in the fifth grade or the eighth grade. This design allows us to model a student's change in academic performance from Grade 3 to Grade 8 as the difference in the two test scores.

For a given grade level, the scale score is associated with a broader indicator of achievement: the *performance levels*. There are three performance levels relevant to our analysis: Below Standards, Meets Standards, and Exceeds Standards²⁷. The performance levels place student achievement in the context of the annual learning standards. For example, if a fifth grader and an eighth grader got a score of 230 on their respective exams,

²⁷ Within the Below Standards designation, defined above, the state characterizes a fourth level of performance ("Academic Warning"), which describes extremely low test scores (e.g., in the bottom ten percent of statewide performance). The distinction is of limited practical relevance in this analysis; excluding students with learning disabilities, the proportion of Oak Park students that fall below this threshold is less than one percent. (ISBE District Report Card, 2006)

they would be considered equally proficient, in terms of the ISAT. However, the performance levels indicate that the fifth-grader had met the learning standards for his grade, whereas the eighth grader did not.

TEST*	Below Standards	Meets Standards	Exceeds Standards	Maximum
Math3	120	184	224	276
Math5	149	214	271	286
Math8	189	246	288	333

*Math3 = ISAT Grade 3 Math Exam

Figure 2 : ISAT Score Ranges and Performance Levels by Grade²⁸

4.2.2. Scale Recalibration

As a barometer of school effectiveness, performance levels can be more informative than scale scores because they place student achievement in the specific context of improving skills related to the learning standards. However, as mentioned in the previous chapter, performance levels require us to organize achievement data into broad categories, an approach which can which can conceal valuable information about differences in group achievement.

Given the distinct advantages of performance levels and scale scores, an appealing compromise would permit us to characterize performance in a way that acknowledges the learning standards, while preserving the analytic benefits of a continuous scale. To achieve this, we introduce a recalibrated scoring scale that allows us to express the scale score in terms of the performance levels.

Figure 3 depicts the recalibrated scale, as applied to the ISAT exam for 8th Grade Math. The extremes of our recalibrated scale are -1 and 1, which correspond to the minimum and

²⁸ The minimum and maximum scores refer to students who took the exam prior to 2006; for students taking the exam in later years, the only “fixed” extreme score is the Grade 3 minimum.

maximum scale scores for a given exam. A score of zero on our scale corresponds to the middle of the “Meets Standards” range, which equals 266.5 for the 8th grade Math ISAT. Our scale is symmetric with respect to the Meets Standards range, with -0.15 and 0.15 corresponding to the lower and upper bounds of the Meets Standards range²⁹. Within performance levels, increasing exam scores are assumed to be proportional to greater accrual of skills, resulting in a piecewise linear relationship between the existing scale and our recalibrated one.

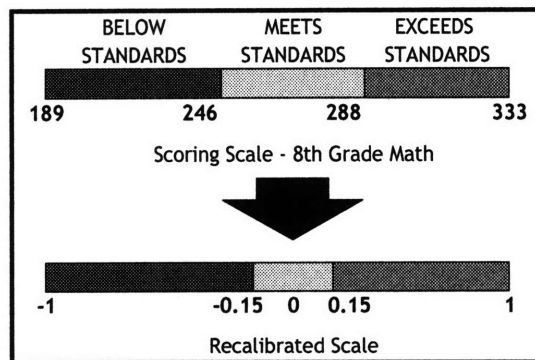


Figure 3: Recalibration of the Grade 8 Math ISAT

With our new process for quantifying score differences in the context of learning standards, we will calculate our absolute gap measures using our recalibrated scale. We compute the *recalibrated mean gap*, *recalibrated median gap*, and the *recalibrated quartile gap* as before, by finding the difference in the (recalibrated) statistics. We calculate absolute metrics by dividing the measured gap by the maximum possible value of that gap. The recalibrated scale takes values from -1 to 1, thus the length of the recalibrated scale is two.

For example, to calculate the recalibrated mean gap between the minority and majority groups, we would use the following equation:

²⁹ The threshold value of 0.15 represents an early hypothesis that the scores which “Meets Standards” might occupy 15% of a given scoring range.

$$MeanGap = \frac{\mu_{RMIN} - \mu_{RMAJ}}{2}$$

In the equation, μ_{RMIN} and μ_{RMAJ} denote the recalibrated group means for the minority and majority groups, respectively. By construction, negative values of the mean gap indicate that minorities are at a disadvantage.

4.3. GRADE 8 TEST PERFORMANCE (2005)

We begin our analysis of the Oak Park achievement gap with a look at the 8th grade class of 2005³⁰. In this section, we compare student performance on the ISAT Mathematics exam. The intent of our study is to characterize achievement trends within the community of Oak Park; in order to mitigate the influence of students who transfer in and out of the district, we will focus on students who have been in the district for a minimum of five years. Occasionally, we refer to these students as “veteran” or “cohort” students to indicate their status as long-term members of the district.

4.3.1. Performance Summary

Before delving into our full collection of achievement metrics, we briefly summarize the eighth grade ISAT results (

Figure). There are 369 records in our dataset; 264 (72%) of the students are in the majority (MAJ), with the remaining 105 in the minority (MIN) group. As the figure shows, the raw scores indicate the presence of an achievement gap

As mentioned in the previous chapter, a commonly cited measure of performance is the “Meets/Exceeds Percentage.” (70%) of the minority eighth-graders in our 2005 cohort met or exceeded standards, but that proportion is noticeably lower than the 95 percent of majority students who were able to meet or exceed the standards.

³⁰ Unless otherwise noted, we identify academic years by the ending calendar year; in this case, we are referring to the 2004-2005 academic year.

The next measures in the table derive from the scale score. These figures also reflect the discrepancy in group performance; on a scoring scale with a range from 189 to 333 (144 units), minority scores were, on average, roughly forty points lower than non-minorities.

8th Grade Math ISAT (2005 Cohort)		
	MAJ	MIN
Pct. Meet/Exceeds	95%	70%
Average	263	300
Median	259	300
Recal. Average	-0.02	0.45
Recal. Median	-0.06	0.38

Figure 4: Overview of ISAT Performance (2005 Cohort)

The third set of measures in the table compares performance using our recalibrated scale, which was mentioned in the previous section. To recall the approach, the student’s scale scores are transformed to a $[-1, 1]$ scale, and a score of zero on the recalibrated scale corresponds to the center of the “Meets Standards” range. Although most of the minority students met the state standards, the recalibrated scores indicate that the middle of the minority score distribution lies on the lower side of the meets standards range, albeit barely. In contrast, we see that the center of the majority distribution is well above the “Meets Standards” upper threshold of 0.15, into the “Exceeds Standards” range of the scale.

4.3.2. Measuring the Test Performance Gap

The summary data leave little doubt of the existence of an achievement gap, but as we have said before, the magnitude of the gap depends on the choice of measurement. In Figure 5, we measure the 2005 8th grade testing gap using the relative and absolute measures introduced in Chapter 3. Although the magnitudes of the metrics vary, we immediately note that all indicators have a negative sign, signaling that the minority group is at a disadvantage.

Because all of our gap metrics are based on a [-1,1] scale, Figure 5 indicates that the discrepancy in relative measures is larger than in absolute measures. In evaluating this outcome, it is helpful to remember that relative measures address the question of “*who’s ahead*”, whereas absolute metrics measure “*who’s ahead, and by how much.*” Together, the metrics tell us that, for these students, the divergence in group performance may be greater than the scoring scale might indicate. On the statewide scale, which goes from 189 to 333, a group average of 300 is already noticeably higher than 263; yet the discrepancy in group performance is amplified when we adapt our focus to the distribution of performance within the Oak Park community. This outcome emphasizes the point that, even within a district where a relatively high percentage of students meet the state standards, substantial achievement gaps can still exist.

ISAT Math Gap, Grade 8 (2005 Cohort)	
	2005 Math8
Students Tested	369
Min. Students	105
Minority %	28%
RELATIVE MEASURES	
Rank Sum Index	-0.62
K-S Index	-0.49
Median-as-Percentile	-0.83
ABSOLUTE MEASURES	
Mean Gap	-0.23
Median Gap	-0.22
Quartile Gap	-0.28

Figure 5: ISAT Gap Metrics for 8th Grade Math (2005 Cohort)

The magnitudes of the relative measures suggest that there is very little overlap in the performance of majority and minority students. As we discussed in Chapter 3, the rank-sum index can be used to approximate the probability that a minority student selected at

random, will outperform a random student from the majority group. Applying the formula $\rho = (RSI + 1) / 2$, a rank-sum index of -0.61 indicates that the likelihood that a randomly-chosen minority student outscores a majority student is about 20 percent. The median-as-percentile score (-0.83) tells us that although half of the minority group scored below 269 (the minority median), only 9 percent of the majority group had scores that low.

The absolute gap metrics are fairly consistent in magnitude, as the three measures occupy a fairly small range (-0.22 to -0.28, on a scale of length 2). The mean and median gaps are nearly identical, indicating that influence of outliers on our measurement is minimal. The quartile gap is slightly higher than the mean and median gap, which is an indication that the group distributions may have a different shape; however, the primary indication is that all three absolute gaps are of comparable magnitude.

As we noted before, the absolute gaps are smaller in magnitude than the relative gaps; in this case, the discrepancy is due to the fact that the nominal difference in group averages understates the lack of overlap in the group distributions. However, this is not always the case; in some scenarios, absolute gaps can be just as large as relative gaps, or even larger.

Our analysis of eighth grade test scores leaves little doubt of an achievement gap on the 2005 ISAT exam. ISAT scores are the most readily available indicator of student skill in Oak Park, and in the following chapters, we will use ISAT performance data to study a variety of questions concerning the nature of the Oak Park achievement gap. But before we get there, we think it important to contrast student performance on the eighth grade ISAT with another important, indicator of student performance.

4.4 GRADE 8 CLASSROOM PERFORMANCE (2005)

Standardized test scores are a commonly accepted measure of student achievement. However, they are not the only measure of student achievement. In addition to ISAT

scores, the district also has a record of the report card grades for the class of 2005. In what follows, we reconsider the Oak Park achievement gap, using student grade data.

4.4.1. Overview

Although standardized tests like the ISAT have become prevalent in achievement gap studies, the tests are not without their detractors. Some researchers believe the standardized tests, in certain circumstances, can overstate proficiency gaps. For example, a long standing criticism of standardized tests suggests that test content reflects the cultural biases of those who create the tests, ostensibly placing minorities at an unfair disadvantage. Although the question of culturally biased testing is outside of the scope of this analysis, the opportunity for such bias on a Math exam might well be substantially lower than, say, a Reading exam.

Cultural bias aside, another criticism of test score analysis stems from the “high-stakes” nature of many exams. The ISAT is the statewide standard for measuring student achievement, and, in keeping with the tenets of the federal No Child Left Behind Act, ISAT results are the primary barometer of school quality. Some researchers argue that importance attached to high-stakes tests like the ISAT fosters a condition known as *stereotype threat* within minority groups³¹. Stereotype threat is relevant to our work because its presence could inhibit test performance, potentially creating a gap attributable to anxiety rather than skill.

Stereotype threat is stress associated with low expectations; the theory suggests that if a student feels that he or she is expected to under-perform on an exam (e.g., because of race), then the associated loss of self-confidence interferes with the student’s actual skill to perform well on the test, leading to lower performance. Although the presence of stereotype threat is not limited to standardized tests, there is evidence that suggests that the

³¹ Steele and Aronson (1995)

effect increases when academic tasks have diagnostic implications, e.g., as in a high-stakes test³².

The imperfect nature of standardized tests suggests that, we should consider alternative measures of student proficiency. Student grades are valuable to our analysis because they reflect aspects of the educational experience that standardized tests are less likely to capture. Specifically, the course grade represents the instructor's view of a student's subject mastery. Whereas the test score from the state exam relies on a one-day assessment of student skills, course grades reflect a continuous body of student work accumulated over several weeks, or months. In Oak Park, this assessment is developed independently of a student's performance on the state test, meaning that course grade provide a complementary perspective on student achievement.

4.4.2. Methods

Schools in Oak Park operate on a trimester system, with course grades and report cards distributed three times a year. The trimester grades are non-cumulative, so the three grades that a student receives are, in theory, independent. As such, we take the simple average of the three grades as the teacher's measure of student performance for that course. We will refer to this value as the student's average Math grade, or *AMG*.

On the report card, students receive one of five letter grades (A, B, C, D, or U). According to the district, course grades reflect skill (as opposed to effort); the "U" grade represents unsatisfactory performance and is equivalent to failing the course for that trimester. The district also gives partial grades, which are denoted by a plus sign (+) or a minus sign (-).

For purposes of calculation, we assume a four-point scale for which the highest letter grade ("A") is worth four points; continuing down the scale, a "B" is worth three points, a "C" is worth two points, and so on. To accommodate partial grades, we add one-third of a point

³² *ibid*

for “plus” grades, and subtract one-third of a point for “minus” grades. As an example, if a student received an A+, an A, and a B+, his AMG for the course would equal $(4.33 + 4.00 + 3.33)/3 = 3.89$.

An overview of the 8th grade report card data for the 2005 cohort is provided in Figure 6. There are fewer grade records than there were for the ISAT exam (317 vs. 369); this is because the “missing” students from this analysis took supplemental or otherwise non-traditional Math courses. The data in this analysis comes from students enrolled in either the “Standard” or “Honors” Math course. In this section, we pool the grade data from both courses, under the assumption that course grades are comparable across courses³³.

For the minority and majority group, we present mean and median AMG, as well as the percentage of students with an average Math grade of C (2.0) or better. The data indicate that there is an achievement gap in the classroom as well. On average, minority students have an AMG of 2.3, which is roughly equivalent to a “C+” average in Math. In contrast, students in the majority group had mean and median AMGs of 3.0 and 3.2, nearly a full grade higher than the minority group. Accordingly, 87% of majority students average a grade of C or higher in the 8th grade Math class, as opposed to two-thirds of the minority group.

8th Grade Avg. Math Grades (2005 cohort)		
	MAJ	MIN
No. of Students	234	83
Mean	3.0	2.3
Median	3.2	2.3
C or better (%)	87%	66%

Figure 6: Overview of 8th Grade AMG (2005 Cohort)

³³ We relax this assumption in our discussion of course placements in Chapter 7.

4.4.3. Findings

The comparison of Math performance in the classroom indicates that there was a gap in the Math grades as well on the state tests. But is the classroom performance gap as large as the testing gap? We address this question by calculating our measures of absolute and relative gaps using the grade data, and comparing those results to the results we get when using the ISAT data (Figure 7).

8th Grade Gap Metrics, Grades vs. Testing, 2005		
	AMG	ISAT
No. of Students**	317	317
Minority Students	83	83
Minority %	26%	26%
RELATIVE MEASURES		
Rank Sum Index	-0.45	-0.57
K-S Index (Rank)	-0.37	-0.46
Median-as-Percentile	-0.53	-0.80
ABSOLUTE MEASURES		
Mean Gap	-0.17	-0.20
Renorm Median Gap	-0.21	-0.20
Quartile Gap	-0.20	-0.25

Figure 7: Classroom and Testing Gaps, 8th Grade, 2005 Cohort

Before discussing the course grade metrics, we recognize that the range of course grades is more limited than that the ISAT score range, thus increasing the likelihood of ties. Our method of handling ties is relevant because it influences our computation of the rank-sum index. To briefly review our approach, we assign average ranks in the case of ties: for example, in the event that there are two students with the lowest AMG, then both students

would receive an average rank of 1.5 (i.e., the sum of the two lowest ranks {1, 2}, divided by 2).

When compared side to side, the classroom and testing gaps are somewhat similar. On the classroom side, again the relative metrics are larger than absolute metrics. However, when compared to the ISAT results, the relative and absolute gaps appear smaller when measured with grade data. For the absolute metrics, the classroom and testing gaps are actually very similar in magnitude, with all metrics occupying a rather narrow range (-0.17 to -0.25) on the [-1,1] scale.

In contrast, the difference between the classroom and testing gaps are considerably larger among the relative metrics, which suggests that there is more overlap in the course grades of than there is in test scores across the two groups. This decrease may be due, in part, to the fact that the range of course grades is smaller than the range of test scores. However, this discrepancy would also support a hypothesis of stereotype threat, as the high-stakes nature of the ISAT may induce higher levels of stress than the day-to-day classroom environment.

4.5. SUMMARY

In this chapter, we began our exploration of race and achievement in Oak Park. Using 8th grade data from the 2005 academic year, we have compared the Mathematics performance of majority and minority students in two major ways, beginning with an analysis of performance on the Math portion of the ISAT and moving on to grades in individual courses. The results for the 8th grade — the last year of elementary school — suggest that fears about an achievement gap in Mathematics were not ill-founded. Relative and absolute indicators about performance gaps were all negative and substantial, indicating that the minority students did not do as well as their White and Asian counterparts.

In an examination of report card grades, we found that achievement gaps were present in the classroom performance as well. On average, Math grades for minority students are a full letter grade lower than average grades for the majority groups. When we compared Math grades using the absolute and relative gap metrics, we saw the same trends as we had with the ISAT data, albeit in smaller magnitudes.

In the next chapter, we expand our analysis to develop a more comprehensive picture of the achievement gaps in Oak Park. Using test data spanning a total of eight years, we develop a sense of how the shape of the achievement gap changes within and across different cohorts of Oak Park students.

5. Evolution of the Achievement Gap

In this chapter, we study the evolution of Oak Park's achievement gap over time. We have seen considerable gaps in the Mathematics achievement of eighth-graders from the class of 2005, but have the 8th grade gaps always been this large? In this chapter, we compare the results from the 2005 cohort to achievement gaps recorded in other cohorts. For these students, we will also compare 8th grade outcomes to test data from earlier grades to understand how achievement trends evolve among students. In addition to macroscopic analysis of changes in student achievement through elementary school, we also study changes in achievement among students of comparable skill.

5.1. THE ACHIEVEMENT GAP ACROSS COHORTS

Previously, we used ISAT data and course grades to explore the 8th grade Math gap among the 2005 cohort. Now, we expand our analysis to the 2004 and 2006 cohorts as well. For all three cohorts, we have ISAT Math results from 3rd Grade, 5th Grade, and 8th Grade. Collectively, these data provide the basis for our further exploration of achievement gaps in Oak Park. As before, we will focus on "veteran" Oak Park students who attended school there from grades 3 through 8

5.1.1. Comparing the Cohorts

Given our knowledge of performance gaps among the 2005 cohort, we measure the eighth grade testing gaps amongst the 2004 and 2006 cohorts for comparison. If the three cohorts show similar performance patterns, then we can pool the data, creating larger sample sizes from which to draw inferences. However, if there are significant differences in the achievement data across cohorts, then pooling the data may conceal important

distinctions across the cohorts. Such distinctions could raise questions about whether any developments in Oak Park might be increasing or decreasing the achievement gap over time.

To the question of similarity across cohorts, a direct comparison of gap metrics (Figure) reveals some similarities, yet remains somewhat inconclusive. For example, although the proportion of “veteran” minority students had increased over time, the eighth – grade test performance of minority and majority students, as measured in scale scores, is fairly consistent across cohorts. However, the rank-sum metric and the median gap metric are larger for the 2005 cohort than for the other cohorts. Also, the Meets/Exceeds metric is a few percentage points higher for minorities in the 2006 cohort.

Eighth Grade Math ISAT by Cohort			
	2004	2005	2006
	Math8	Math8	Math8
Students Tested	363	369	373
Min. Students	98	105	119
Minority %	27%	28%	32%
Median (Min)	264	259	261
Median (Maj)	297	300	297
Rank Sum Index	-0.54	-0.62	-0.58
Median Gap (Recal.)	-0.17	-0.22	-0.18
Pct. Meet/Exceeds (Min)	70%	70%	73%
Pct. Meet/Exceeds (Maj)	95%	95%	95%

Figure 1: 8th Grade Gap Metrics for Math ISAT (2004-06 Cohorts)

The achievement data reveal similar behavior across the cohorts, but there is some variation in the data. The question of whether the variations represent statistically *significant* differences across cohort remains unresolved. Rather than pool the cohorts, and

move immediately into studying the evolution of the gap, we will digress briefly to outline a formal approach for testing the level of similarity across the cohorts.

5.1.2. Testing Similarity across Cohorts

5.1.2.1. The Rank-Sum Statistical Test

Our mechanism for testing for similarity across cohorts is the rank-sum statistical test³⁴. (This test is not to be confused with the Rank Sum Index, which we use to a relative measure of achievement gaps. Given an initial (null) hypothesis (e.g., that two cohorts share a common underlying distribution, and that the observed differences reflect a form of sampling error³⁵), the rank-sum test assesses the degree of deviation across datasets, with the assumption that the hypothesis is correct. The output statistic of the rank-sum test, referred to as p , equals the likelihood of observing at least that level of deviation, given that the hypothesis is true. In other words, the test says something about the plausibility of the hypothesis, in light of the data.

For this analysis, we test the following null hypothesis:

H₀: The severity of the achievement gap does not vary significantly across cohorts.

Now, although the rank-sum test makes comparisons among the distributions of two datasets, we are interested in comparing three cohorts of data. We address this issue by applying the rank-sum test twice. As an example, suppose the relevant data for the three cohorts are labeled Set A, Set B, and Set C. Then, in conducting a two stage test, we replace the original null hypothesis with two hypotheses:

³⁴ The rank-sum statistical test is not to be confused with the Rank Sum Index, which we use to measure (relative) achievement gaps.

³⁵ The tacit assumption is that (say) the 119 minority group students in the 2006 cohort are a random sample from all Oak Park minority group students over a longer period, the performance of which is stable over time.

H0': The severity of the minority achievement gap does not vary significantly across Sets B and C.

H0'': The severity of the achievement gap does not vary significantly across Set A and the combined Set (B+C).

If both of these hypotheses pass the rank-sum test, then we can accept our original hypothesis and group the data

5.1.2.2. Results

In order to apply the Rank Sum test, we need to represent the data for each cohort in a way that depicts the racial gap among the students of that cohort. Since we are interested in comparing absolute and relative achievement gaps, we developed a representation for both types of gaps.

We summarize the results of the two-stage rank-sum tests in Figure 2; in short, we fail to reject the null hypotheses and consider the behavior of the three cohorts to be close enough to combine the data. In all, we tested four hypotheses:

- 1) *The relative achievement gap does not vary significantly across the 2005 and 2006 cohorts³⁶.*
- 2) *The relative achievement gap does not vary significantly across the 2004 cohort and the combined 2005 and 2006 cohorts.*
- 3) *The absolute achievement gap does not vary significantly across the 2005 and 2006 cohorts.*
- 4) *The absolute achievement gap does not vary significantly across the 2004 cohort and the combined 2005 and 2006 cohorts.*

In our test, had any one of the four hypotheses failed, we would have rejected the claim of similar achievement gaps across cohorts. The decision to reject is based on the *p*-value; in

³⁶ The order in which we combine cohorts can affect the results of the rank-sum tests, but in all cases, our conclusion (i.e., that the cohorts can be combined) remains the same.

a single hypothesis test, when p is less than 0.05, it is common to reject the hypothesis, because at the point the probability of Type I error (i.e., that we accidentally reject) is only 5 percent. When multiple tests are used in hypothesis testing, the overall likelihood of Type I (false positive) error increases. The Bonferroni correction is one of many approaches to mitigating error; for n tests and an overall Type I error rate of α , the suggested threshold for each test is

$$p^* = 1 - (1 - \alpha)^{1/n}$$

For a test of 4 hypotheses and an overall Type I error of .05, the suggested rejection threshold for each hypothesis is $p = 0.013$. As Figure 2 shows, the p -value of each hypothesis is 0.20 or higher, so we fail to reject any of the hypotheses.

Similarity Test for Cohort Data	
<i>Hypotheses</i>	<i>p*</i>
MIN percentiles	
H1) 2005 vs. 2006	0.81
H2) 2004 vs. (2005 & 2006)	0.69
Deviations from MAJ Avg.	
H3) 2005 vs. 2006	0.20
H4) 2004 vs. (2005 & 2006)	0.26
* p equals the likelihood of outcomes, assuming the groups have a shared distribution	

Figure 2: Comparison of Cohorts using Rank-Sum Testing

Testing indicates the three cohorts essentially exhibit the same behavior, meaning that the district achievement gaps have been stable over the three year period. This finding was consistent with the district's own assessment, that there had no evidence to suggest that the nature of the gap had changed appreciably over time. Given the observed constancy of achievement trends, we pool student outcomes from the three cohorts to test earlier hypotheses concerning achievement in the district. Pooling the data provides a larger dataset, which improves our ability to detect performance trends.

5.2 CHANGES IN GROUP ACHIEVEMENT OVER TIME

5.2.1. Comparing Test Gaps across Grade Levels

Given the magnitude and persistence of the district gap in the eighth grade, we assess the magnitude of the achievement gap in earlier years. All of the students in our sample took the ISAT math exam in Grades 3, 5, and 8. By comparing the eighth grade results to the data from earlier grades, we develop a sense for how the gap has evolved as the students got older.

The ISAT results and gap metrics for all three examinations (Figure 3) collectively indicate that the difference in the minority and majority group performance has neither grown nor shrunk in any consistent way over time. On the contrary, the achievement gap has been fairly consistent from Grades 3 through 8. Relative measures of the gap, such as the rank-sum index, show a negligible amount of change in the gap. The variation in absolute gap metrics, such as the recalibrated median, is also quite small (deviation of 0.11 units on a scale of length 2). With respect to performance levels, the *Meets/Exceeds* data show nearly identical patterns in achievement, particularly in the first (Grade 3) and final (Grade 8) state exams.

ISAT Overview, Grades 3, 5, 8			
	ISAT3	ISAT5	ISAT8
Median (Min)	199	224	262
Median (Maj)	231	257	299
Rank Sum Index	-0.58	-0.59	-0.58
Recal. Median Gap	-0.15	-0.09	-0.20
Pct. Meet/Exceeds (Min)	72%	67%	72%
Pct. Meet/Exceeds (Maj)	96%	95%	95%

Figure 3: ISAT Overview Data (2004-2006 Cohorts)

Although the exam content varies across grade levels, all exams are linked to a common interpretation of scores. For example, a student who gets a score of 230 on the third grade exam is assumed to possess the same abilities as a student who attains a 230 on the fifth-grade exam. This feature permits the direct comparison of scale scores across grade levels, adding additional perspective to the district gap. For example, note that the median score for majority students in the third grade is higher than the median score for minorities in the fifth grade. Further, the median score for majority students in the fifth grade is nearly equal to the median score for minorities in the eighth grade. In this view, minority students, as a group, are roughly 2-3 years behind the majority group counterparts for most of the time they are in the district.

5.2.2. Pre- ISAT Achievement Gaps

In addition to the ISAT exams, provided by the state, the students in our study were also required to take the Stanford Achievement Test, 9th Edition (SAT). The first ISAT exams are administered during the spring of a student's third-grade year. In contrast, these students took the SAT in the fall of their second-grade year, and again in the fall of the third-grade year. Thus, the students in our study took two Stanford exams before they began their state-mandated testing.

Since our earliest measures of achievement come from the Stanford data, we compare the gaps in the Stanford data to our findings from the ISAT data. This comparison is justified because the ISAT and the Stanford test are comparable in format and rigor³⁷. Scoring scales for the Stanford and ISAT exam are independent of each other, and employ different methods of calibration, so direct comparisons of scale score data are not useful. However, we can compare results from the two exams with the absolute and relative metrics developed in **Chapter 3**.

³⁷ *Illinois State Assessment Technical Manual* (1999 version)

A side-by-side comparison of relative and absolute gap metrics³⁸ for the two sets of exams (Figure 6) supports the notion that the achievement gap attains its magnitude early, and persists as the student body gets older. The first two columns present the relative gap metrics from the Stanford test. There is a nominal decrease in the gap metrics between the second and third grade Stanford exams, yet the differences are minor compared to the magnitude of the metrics.

Achievement Gap Metrics, Grade 2 - 8 (2004-2006 Cohorts)					
	2004-6 Stan2	2004-6 Stan3	2004-6 Math3	2004-6 Math5	2004-6 Math8
	<i>STAN2</i>	<i>STAN3</i>	<i>ISAT3</i>	<i>ISAT5</i>	<i>ISAT8</i>
RELATIVE MEASURES					
Rank Sum Index	-0.50	-0.45	-0.58	-0.59	-0.58
K-S Index (Rank)	-0.39	-0.36	-0.44	-0.45	-0.46
Median-as-Percentile	-0.66	-0.59	-0.75	-0.76	-0.79
ABSOLUTE MEASURES*					
Mean Gap	-0.12	-0.09	-0.17	-0.16	-0.20
Median Gap	-0.11	-0.08	-0.15	-0.09	-0.20
Quartile Gap	-0.13	-0.09	-0.17	-0.14	-0.24
*Stanford 9 score ranges for absolute metrics estimated from district data					

Figure 4: Relative Achievement Gap Metrics, 2004-2006 Cohorts

Interestingly, the gap measures show their largest deviations between the Stanford third-grade exam, given in autumn, and the Grade 3 ISAT, given during the following spring. In terms of magnitude, it is surprising that the gap metrics would change so noticeably in a matter of months. To examine the difference further, we compare Grade 3 ISAT scores among students who had similar performance on the Grade 3 Stanford exam.

³⁸ For the Stanford absolute metrics, score ranges are estimated from district data.

In summary, we group the students according to their Stanford exam score, and in each of the deciles, we compare the average ISAT scores for minority and majority students (Figure 5). As one would expect, there is clear correlation between Stanford and ISAT exam scores; average ISAT scores for both groups increase as we move across deciles of increasing performance on the Stanford test. However, the data show that, within deciles, minority students routinely obtain lower scores on the ISAT than their similarly situated peers in the majority group.

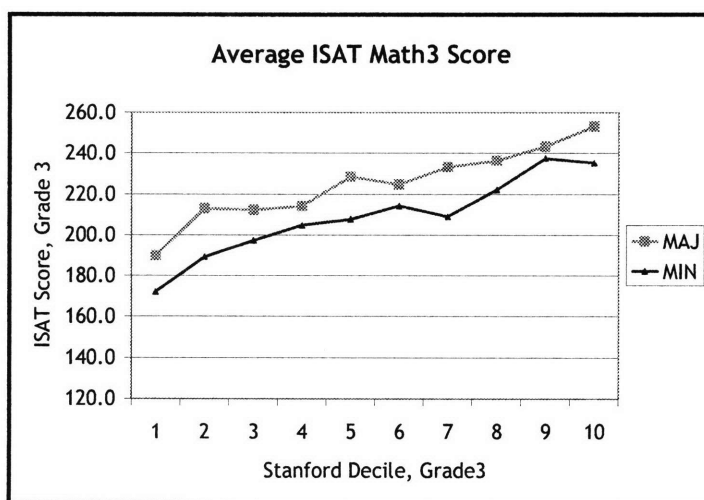


Figure 5: Average Grade 3 ISAT (Spring) by Grade 3 Stanford Decile (Autumn)

Although we do not know for sure, we reason that the growth in the third-grade metrics could be due either to differences in exam content, differences in the exam context, or both. The first explanation seems unlikely; although differences in content might make the ISAT exam more capable at discerning differences in achievement, this capability would not explain the distinct advantage shown by the majority group. Recall that the Stanford exam is an in-school diagnostic, whereas the ISAT is a high-stakes state exam; if we allow for the possibility that two exams might be perceived differently by the student, then the trend might indicate the presence of stereotype threat among minority students aware of the relative importance of the ISAT exam.

5.3. CHANGES IN STUDENT PROGRESS OVER TIME

5.3.1. Group Achievement vs. Student Progress

As we compare the behavior of the minority group and the majority group, the data suggest that differences in group performance are largely consistent across several grade levels starting at grade 2. Although we do not have the data, it seems plausible that there was a readiness gap between these students before they began school. As noted in **Chapter 2**, a central question regarding the achievement gap concerns the ability of schools to counteract the socioeconomic and environmental factors believed to hinder minority achievement. Judging from analyses of these groups, the relationship of school effects to increases or decreases in the achievement gap appears to be minimal.

Given the apparently fixed position of the overall achievement gap, what can we say about the achievement of individual students in Oak Park? Surely, individual changes in performance are not fixed; students enter the Oak Park district at varying levels of skill, and make various amounts of progress while they are in school. It is possible that the constancy in aggregate is actually the sum of divergent patterns among different groups of students? Beyond comparisons of aggregate behavior, we move towards comparing changes in progress among students who exhibit similar levels of skill.

To classify students according to “initial skill”, we will use the performance levels from the 3rd Grade Math ISAT. Although we have taken issue with the use of ISAT performance levels as a measure of achievement gaps (c.f. **Chapter 3.1**), we believe that performance levels provide a reasonable, if imperfect, choice for grouping students of similar skill³⁹. In keeping with the names of the performance levels (Below, Meets, and Exceeds Standards), we occasionally use the terms *Below3*, *Meets3*, and *Exceeds3* to refer to the students with the given classification.

³⁹ Hanushek and Rivkin (2006) were less enthusiastic about using prior math performance for classification, citing the potential for skewed results stemming from regression to the mean.

5.3.2. Transition Frequencies

The distributions of third-grade performance levels for both ethnic groups are presented in Figure 6. As the figure shows, most minority students are in the *Meets3* group, whereas most of the majority group is in the *Exceeds3* group. If a student demonstrated a certain level of performance in the 3rd grade, how likely were they to raise or lower that level of performance over time? In what follows, we use historical data to categorize the various performance trends among our students; these transition frequencies allow us to summarize changes in performance and also provide a basis for comparing differences in progress across ethnic groups

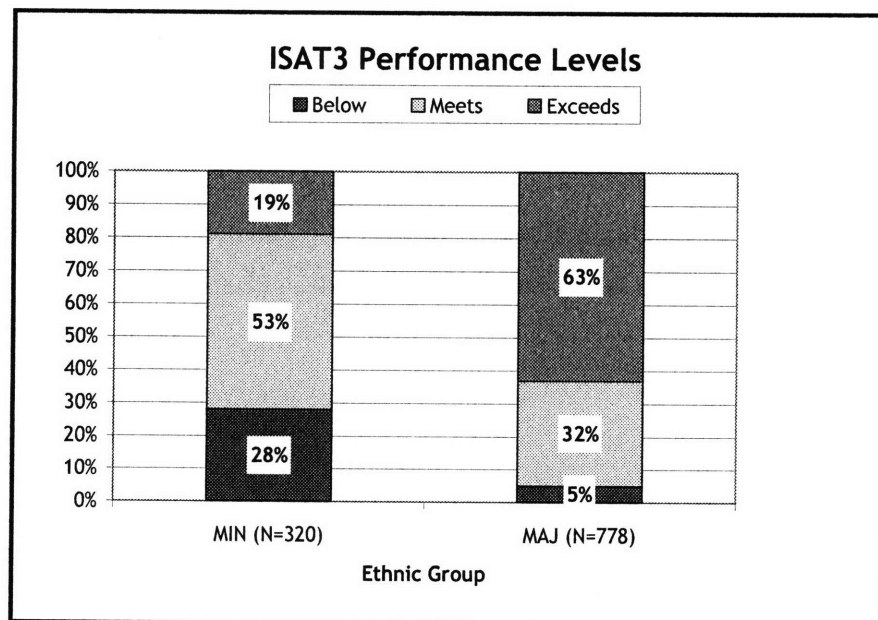


Figure 6: Grade 3 Performance Levels by Ethnic Group

We define student transitions by comparing the 3rd Grade performance level with the 8th Grade performance level. For example, a student who has a performance level of *Meets Standards* on the 3rd Grade exam will have an 8th Grade level which is either higher (*Exceeds Standards*), lower (*Below Standards*), or the same (*Meets Standards*) as the

earlier rating. There are three possible transitions from the Meets Standards level, and two possible transitions from the Below Standards and Exceeds Standards levels, yielding a total of seven transitions to consider (Figure 7).

Grade 3 Performance	Grade 8 Performance
Exceeds Standards	Same
	Lower
Meets Standards	Higher
	Same
	Lower
Below Standards	Higher
	Same

Figure 7: Grade 3- Grade 8 Performance Transitions

5.3.2.1. Transition Frequencies in Oak Park

Using these categories, we calculate transition probabilities for the students in our sample (Figure 8). The transition probabilities represent the actual outcomes of Oak Park students, and they allow us to model the likelihood that a student’s performance level will improve, decline, or remain the same over time, given their Grade 3 performance. The transitions are mutually exclusive and sum to one for each of the performance level subgroups (Exceeds3, Meets3, and Below3).

In general, the results indicate that the students in Oak Park will either maintain or improve their performance level over time. Students who exceed standards early show the strongest retention rate of all performance groups, with 80 percent of them exceeding standards at the 8th grade level as well. Among students who meet standards early, most students (62%) match their earlier performance when they take the 8th grade exam, but among those whose did not, Oak Park students were more than twice as likely to improve their level of performance rather than decline (26% vs. 12%). Although most students below standards in the third grade remained below standards in the eighth grade, 40% of these students were able to meet the eighth grade standards, which is a positive sign for the district.

The transition frequencies help explain the consistency we saw in the third grade and eighth grade achievement gap metrics. Across all levels of performance, over seventy percent of Oak Park students experience no net change in their performance level in grades 3 through 8. Among the remaining students, the percentage of students who increase their performance level was nearly identical to the percentage of students whose performance level declined.

Grade 3 Performance	Grade 8 Performance	Frequency
Exceeds Standards	Same	80%
	Lower	20%
Meets Standards	Higher	26%
	Same	62%
	Lower	12%
Below Standards	Higher	40%
	Same	60%
All Students	Higher	14%
	Same	71%
	Lower	15%

Figure 8: Grade 3- Grade 8 Transition Frequencies

5.3.2.2. Transition Frequencies by Ethnic Group

The transition frequencies describe performance trends throughout long-term Oak Park students, conditioned on third grade performance. In order to determine whether these trends vary with race, we calculate the transition frequencies separately for majority and minority students. If ethnicity is unrelated to student transition frequencies, we would expect the transition frequencies for majority and minority students to be about the same, allowing for some variation due to chance. (Of course, the proportion of students at each starting level can vary between the groups.)

A comparison of transition frequencies for both ethnic groups (Figure 9) reveals that minority students were less successful than majority students at either raising their performance over time, maintaining earlier levels of performance, or both. The difference in outcomes is most apparent among the students who initially exceeded standards. Among students in this group, majority students were far more likely than minorities to exceed standards again on the eighth grade exam (83% vs. 57%).

Among students meeting 3rd grade standards, the likelihood of meeting standards on the eighth grade test was about the same for both groups (61% vs. 64%). However, when the performance level did change, majority students showed a strong tendency to improve rather than decline. In contrast, similarly situated minorities were almost equally likely to improve or decline, with a slight edge to the latter. Among students initially below standards, the difference across ethnic groups is not as pronounced, yet minority students again were less likely to improve their level of performance over time.

Grade 3 Performance	Grade 8 Performance	Frequency	
		MAJ	MIN
Exceeds Standards	Same	83%	57%
	Lower	17%	43%
Meets Standards	Higher	33%	16%
	Same	61%	64%
	Lower	7%	20%
Below Standards	Higher	47%	37%
	Same	53%	63%
All Students	Higher	13%	19%
	Same	74%	62%
	Lower	13%	19%

Figure 9: Grade 3 - Grade 8 Transitions by Ethnic Group

The transition frequency data seem to contradict our earlier notions about the Oak Park gap. Earlier in this chapter, when we measured group achievement, we saw almost no change in the magnitude of the achievement gap from Grade 3 to Grade 8. However, among students with comparable third grade performance, minority students were less successful than their counterparts over this period. Therein lies the apparent contradiction: if minorities are less successful over time, then how does the overall gap stay the same?

The main reason why is due to the difference in initial (third-grade) skill level across ethnic groups. As we saw in Figure 6, the minority group mostly comprised of students who were below or meeting the standards in third-grade, whereas most of the majority students exceeding the third-grade standards. In the “Exceeds” category, there is nowhere to go but down if the performance level changes; in the “Below” category, the opposite is the case. For this reason, this difference in starting-level composition makes it more difficult to detect further differences in progress among students of similar skill.

The concealing effect is noticeable in the combined student transitions, located next to the “All Students” heading in Figure 9. When we look at the combined transitions, a weighted average, we see that the frequencies blend into a symmetrical pattern when grouped together. Overall, most students experience no net change in performance and the students who do are equally likely to improve or decline. We saw this symmetry when we looked at the overall transition frequencies for the district, and now we see it again within ethnic groups. As a result, although more than 1 in 4 Oak Park students changed performance levels over time, overall group performance over this period remains largely unchanged. But, as noted, minority students at particular levels do fare worse than majority students at those levels.

5.3.3. Statistical Significance

Although the transition frequencies in Figure 9 favor the majority group, it is important to determine the statistical significance of these findings. Statistical significance increases the

likelihood that our findings reflect actual gaps within the district, rather than fluctuations from measurement error. In a brief digression, we outline our method for evaluating the statistical significance of these transition data.

Our tool for determining statistical significance in this case will be the Chi-Square test. Given a null hypothesis (i.e., an initial assumption about the “true” data relationship), the Chi-Square test tells us the likelihood of the observed outcome. We represent likelihood with the symbol p ; if p falls below a certain threshold, we reject our null hypothesis. For this test, we adopt the following null hypothesis:

H0: The transition frequencies of minority and majority students come from a common distribution.

The null hypothesis reflects the assumption that the combined frequencies in Figure 8 actually apply to both groups, and that the differences observed in Figure 9 are a result of measurement error. To conduct the test, we compare the *actual counts* of students who made each type of transition to the *expected counts*, which we derive from the transition frequencies in Figure 8.

We conduct a total of four Chi-Square tests: one for each of the Grade 3 performance levels, and one more for all students combined. For students in the “Meets Standards” group, there are three possible outcomes (Higher, Same, Lower) and two ethnic groups (MIN, MAJ), yielding a total of six categories. The test for all students also has six categories. Tests for the “Below Standards” and “Above Standards” groups have four categories (2 outcomes \times 2 ethnic groups).

The Chi-Square analyses (Figure 10) confirm that the difference in transition frequencies across ethnic groups is too large to ascribe to random fluctuations in the data. For each test, the table display the overall transition frequencies (*Freq.*), the actual and expected student counts ($a(i)$ and $e(i)$, respectively), and the likelihood the deviations under the null

hypothesis (p). Every test returns a p -value of 0.00, leading us to strongly reject the null hypothesis and conclude that transition frequencies, and therefore, changes in student performance over time, vary significantly with race.

Grade 3	Grade 8	Freq.	MAJ		MIN		p^*
			a(i)	e(i)	a(i)	e(i)	
Exceeds Standards	Same	80%	405	392	33	46	0.00
	Lower	20%	84	97	25	12	
Meets Standards	Higher	26%	82	66	27	43	0.00
	Same	62%	152	155	106	103	
	Lower	12%	17	30	33	20	
Below Standards	Higher	40%	17	13	53	33	0.00
	Same	60%	15	19	31	51	
All Students	Higher	14%	99	111	80	44	0.00
	Same	71%	572	548	170	218	
	Lower	15%	101	114	58	45	

Figure 10: Chi-Square Analysis of Transition Frequencies

5.3.4. A Closer Look at Math3 – Math8 Progress

In our analysis of transition frequencies, we used Grade 3 performance levels to group children of comparable skill. But how close is “comparable”? Performance levels cover a considerable range of test outcomes, and although we believe that it is fair to say that the students are close in skill, the performance levels are too broad to claim that the students within are “equally” skilled. For instance, consider the students within the Grade 3 “Meets Standards” group, which cover a scoring range of 40 points. If the minority students cluster near the bottom of the range, and the majority students cluster near the top of the range, then it would be misleading to suggest that the two groups demonstrate “identical” levels of Grade 3 performance.

To address the potential issue, we conduct a complementary analysis of the changes in student performance. Rather than rely on the three performance levels, this approach focuses on the actual scale score. Our intent is to find groups of minority and majority students with nearly identical third-grade ISAT (ISAT3) scores, and compare the average gains made by both groups. In doing so, our approach was to create the narrowest test intervals possible, provided that:

1. The intervals are of equal size; and,
2. Each interval contains at least one member of both ethnic groups.

Our approach led us to create intervals of width 4, dividing the range of 3rd Grade scale scores into approximately 39 small intervals⁴⁰. Of the 39 intervals, 30 contain at least one student from each ethnic group. Within each of the 30 intervals, we calculate the average gain in scale score (ISAT8 – ISAT3) for minority students, and subtract it from the average gain for the majority group.

The resulting metric is the *difference in average gain* for that interval. The sign of the metric indicates which groups had the larger gain. If the value is positive, then the majority students in that interval had the higher average gain. Conversely, a negative value indicates that the minority group average was larger within that interval. For example, suppose we had a group of students (“Group #10”) who all scored between 160 and 164 on the 3rd Grade ISAT. Within Group #10, the average minority student score increases by 80 points on the Grade 8 exam, and the average majority student score increases by 85 points. Therefore, the difference in average gain for Group #10 is $(80-85) = -5$ scale points.

The comparison of average gains (Figure 11) illustrates the dominance of the majority group in terms of scoring gains. Each interval represents an independent test of which

⁴⁰ The width of the interval can be adjusted according to the desired level of precision.

ethnic group will make the larger average gain, and in 25 of 30 intervals, the students in the majority group made the larger gains.

This result would be very unlikely if math gains and ethnicity were uncorrelated. To conceptualize the degree of dominance, this result would be akin to flipping a fair coin 30 times and observing 25 “heads.” From the binomial distribution, the likelihood of a result this lopsided (or greater) is about 1 in 17,000.

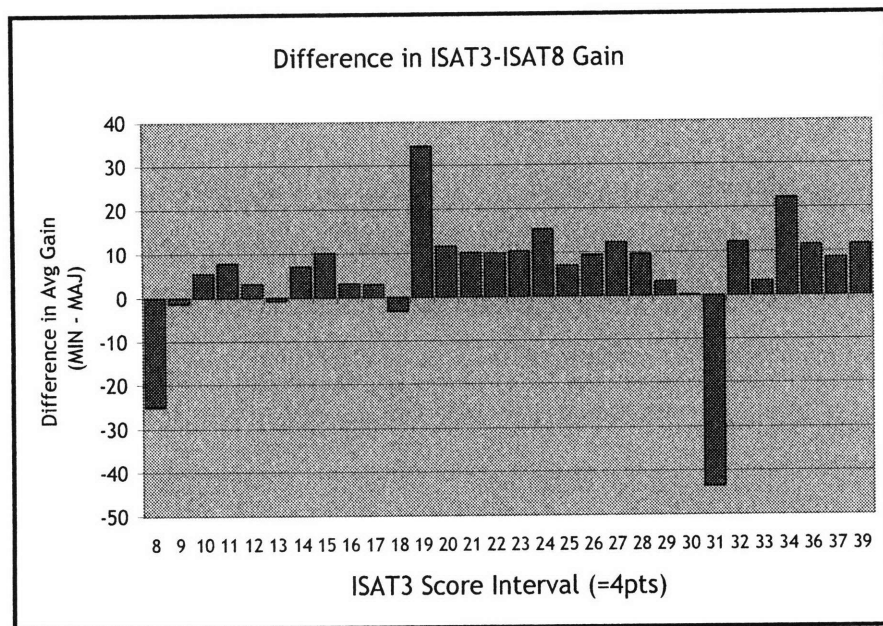


Figure 11: Comparison of Math3 – Math8 Progress with Score Intervals

Performance gains among majority students are more frequent, and also larger in magnitude, as shown by the “margin of victory” within the intervals. For the intervals in which majority students outperform minorities, the difference in average gains is generally ten points or larger. In contrast, of the 5 intervals in which the minority students dominate, 3 of the 5 represents differences of 5 points or less. Further, 4 of the 5 intervals are on the lower half of the scoring scale. This analysis demonstrates that even when we compare

students with nearly equal Grade 3 performance, minorities consistently make less progress than their peers.

5.4. SUMMARY

As we expand our analysis of the Oak Park math gap, the data indicate that the achievement gap has been a constant presence within the district in recent years. First, we found that the patterns we observed in the 2005 cohort did not differ significantly from the patterns we observed in the 2004 and 2006 cohort. This result led us to combine the data from the three cohorts into a single all-inclusive dataset.

Using this expanded dataset, we studied previous test data and found that the eighth grade testing gap is similar in magnitude to testing gaps recorded in earlier grades. Using ISAT data from Grades 3 and 5, and Stanford data from Grades 2 and 3, we found that the relative gap, in particular, remained fairly stable over the years. The evidence suggests that the gap in group achievement is largely constant throughout the years in the district.

An analysis of individual changes in performance over time indicates that the persistence of the gap across grade levels is actually the net result of two dominant achievement trends. Regardless of race, a student's general level of achievement tends not to vary; students at the top of the class tend to remain there over time, and students who struggle early tend to continue to struggle. Also, students who do change their level of performance are almost equally likely to raise or lower their performance, creating a symmetry which results in no net change to group performance.

Surprisingly, despite the dominant indication of no change in the gap, a comparison of performance transitions made by "similarly situated" students provides clear evidence that minorities are less likely than similarly situated students in the majority to improve their performance level over time. In the case of high achieving students, minorities are also less likely to maintain high levels of performance. This seemingly contradictory finding is due

to different distributions of initial performance among the ethnic groups, in addition to the general tendency to remain at a given performance level, regardless of ethnicity.

Given the rather stationary view of group achievement in Oak Park, and the relative lack of progress made by minorities over time, the next two chapters delve into the investigation of other factors that may relate to minority math achievement in Oak Park. The topics covered in the following chapters reflect the interests of district leaders and the availability of data. **Chapter 6** examines the relationship of mobility, economic status, and gender to the Oak Park gap, and in **Chapter 7**, the focus shifts to areas in which the district may be able to better manage the gap.

6. Correlates of Achievement in Oak Park

Through application of alternative metrics to measure achievement gaps and a series of analyses of the gap over time, we have been able to identify some trends regarding the achievement gap faced by the district. In this chapter, we examine the prevalence of other factors believed to influence achievement, namely, differences in student mobility, economic status, and gender. To the extent that these factors may also vary with race, there is the possibility that any one of these factors might contribute to the achievement gap.

6.1. STUDENT MOBILITY

6.1.1. Overview

Over the course of this study, district officials expressed concern over the possibility that differences in student mobility were contributing to the achievement gap. Student mobility refers to the transfer of students in and out of a given school for reasons other than academic promotion. In the district, student mobility was generally believed to be more prevalent among the minority students; under the assumption that mobility hinders academic performance, there was some belief that the race-based gap was due, in part, to increased student mobility among minorities. For example, previous studies have linked student mobility with lower high school graduation rates (Rumberger and Larson, 1998), as well as lower test scores (Engec, 2006).

Our analysis of the Oak Park achievement gap draws primarily from students who were present in the district for eighth grade, and who enrolled in the district for five years or more. We focus on these “veteran” students because their collective performance implies multiple years of exposure to the Oak Park school system, and, as a result, their data

provide the most information with respect to long-term achievement trends within the district. However, in any given year, there are a number of students who transfer in or out of the district. Due to limited data, we can not say as much about the performance of these students, but in what follows, we make some statements regarding their performance.

6.1.2. Mobility Trends in Oak Park

Given our focus on Oak Park’s veteran students, a natural first step is to determine how many students in the district did not meet this requirement. We will refer to these students as relative “newcomers” to the district. Determining the size of this group is relevant because if the number of newcomers is large, than the achievement gap among veterans may vary considerably from the overall achievement gap.

For a given group of eighth-graders, a simple way to determine the size of the newcomer population is to count the total number of students who took the ISAT exam, and compare that number to the number of veteran students. The table in Figure list student counts by academic year and ethnic group. For example, the first two columns of the first row indicate that 326 students in the majority group took the 8th Grade Math Exam in 2004, and of those 326 students, 264 were district veterans. Adding across three years’ worth of data; we see that over 80 percent of all White and Asian 8th-graders had been in the district for at least five years. Over that same period of time, only half of the minority eighth graders had been in the district 5 years earlier.

Test Year	MAJ Group		MIN Group	
	Total	5+ Years	Total	5+ Years
2004	326	264	208	98
2005	326	262	209	104
2006	313	252	223	118
2004-6 Total	965	778	640	320
Pct. w/5+ Years	81%		50%	

Figure 1: Student Counts by Test Year, 2004-2006

The student counts show that, from a district enrollment perspective, minorities are a far less stable lot than the majority group. However, the effect of this difference in stability on the Oak Park achievement gap is unclear. As we have already seen, the gap among veteran Oak Park students is fairly large on its own; if we were to consider the district wide achievement gap among “veterans”, and add in the results of newcomers, could these differences in stability make the gap larger?

In the case of Oak Park, the answer is no. In Figure 2, we compare the 8th grade gap metrics for the all district test-takers to our earlier results for the veterans. From the sample size of the full group, we see that the actual proportion of minority students in Oak Park is much higher than the veteran data would suggest. Although the district is 40% minority at any given point in time, the lower levels of stability among minorities explain why minorities only represent 30% of the veteran group. For both groups, median scores are slightly lower across the full group, which indicates that, within both ethnic groups, veteran students outperform the newcomers.

ISAT Mathematics Gap, Grade 8 (2004 - 2006)		
	Veterans Math8	Full Group Math8
No. of Students	1098	1605
Minority Students	320	640
Minority %	29%	40%
Median (Min)	262	255
Median (Maj)	299	295
Rank Sum Index	-0.58	-0.61
Median Gap	-0.20	-0.18
Pct. Meet/Exceeds (Min)	72%	64%
Pct. Meet/Exceeds (Maj)	95%	93%
Pct. Exceeds (Min)	19%	14%
Pct. Exceeds (Maj)	63%	59%

Figure 2: Gap Metrics for Veteran and Overall Students (2004-2006)

According to our data from the 8th grade exam, veteran students outperform students who arrive later. Also, minority 8th graders are far more likely to have arrived later. Combining these findings, there is an indication that the overall 8th grade gap would be smaller if the minority population were more stable. However, the gap metrics show that the magnitude of the gap of the district wide gap is roughly the same as the gap amongst the veterans. In other words, although student mobility does have predictive power on achievement across the district, the achievement gap is indicative of a long-term trend within the district.

Ultimately, the gap metrics among veteran students refute the notion that high mobility is a leading contributor to the minority Oak Park achievement gap. However, high mobility may limit the district's ability to close their testing gap. If the district were to pursue an educational intervention which improves student achievement over a period of several years, lower stability among minorities suggest that a disproportionately low number of

minority students would be in the district long enough to receive the full benefit of the intervention.

6.2. FAMILY INCOME

6.2.1. Overview

There is a general consensus that a child's academic performance correlates strongly with the socio - economic status of their parents⁴¹. A person's economic status is obviously an indicator of their ability to acquire resources; however, economic status also correlates with one's personal level of education. These relationships support the link between household income and student achievement, for they imply that children from wealthier families will tend to have parents with higher levels of education and broader access to educational resources for their children, such as computers and/or tutoring.

If causal, the link between household income and achievement is relevant because minorities, as a group, tend to earn less than White families. If, on average, minority households in Oak Park earn less than majority households, then the income gap between the two groups may very well contribute to the achievement gap. As a result, we are interested in determining whether these trends exist among the students in our study and what effect these trends may have on the achievement gap in this district.

6.2.2. Low-Income Status

Although Oak Park is, by all accounts, a "middle-class" community, there are differences in income throughout the district. Our primary indicator of economic status is the student's eligibility to receive free or reduced lunch from the district. The state classifies students who qualify for the lunch subsidy as "low-income" students.

⁴¹ Notably, a broad (70+ studies, 100,000 + students) meta-analysis of education studies (Sirin, 2005) found "moderate to high" correlation between socioeconomic status and achievement.

Among the students with at least five years in the district, only 12% qualified for the lunch subsidy (Figure 3). However, within ethnic groups, it is clear that minority students were far more likely to come from low-income households. More than 1 in 3 minority students in our sample are low-income students, whereas fewer than 5 percent of majority students received the subsidy. As a result, the low income population of the school district is composed almost entirely of minority students.

	Both Groups	MIN	MAJ
No. of Students	1098	320	778
Low-Income Students	135	114	21
Low-Income %	12%	36%	3%

Figure 3: Low-Income Students by Ethnic Group (2004-06 Veteran Cohorts)

With such a high concentration of minority students in the “low-income” group, it appears possible that the differences in performance across ethnic groups are more indicative of differences in economics rather than race. In Figure 4, we compare the Grade 8 achievement gap between ethnic groups to the 8th grade achievement gap between subsidized (low-income) and unsubsidized students.

When we calculate the difference in achievement between subsidized and unsubsidized students, we find that the “poverty gap” is just as large, if not larger than, the race-based gap. Raw scores, relative measures, and the Meets/Exceeds Percentage all indicate that the poverty gap is larger than the race gap. In this regard, low-income status appears to be just as powerful a predictor of achievement gaps as ethnic affiliation, inviting the possibility that variations in economic status may be exacerbating the racial gap in achievement.

	c04-06 Math8		c04-06 Math8
No. of Students Tested	1098	No. of Students	1098
Minority Students	320	Low-Income Students	135
Minority %	29%	Low-Income %	12%
Median (Min)	262	Median (LI)	251
Median (Maj)	299	Median (Non-LI)	293
Rank Sum Index	-0.58	Rank Sum Index	-0.63
Median Gap (Recal.)	-0.20	Median Gap (Recal.)	-0.18
Pct. Meet/Exceeds (Min)	72%	Pct. Meet/Exceeds (LI)	63%
Pct. Meet/Exceeds (Maj)	95%	Pct. Meet/Exceeds (Non-LI)	92%

Figure 4: District Achievement Gaps, Based on Race and Low-Income (Reduced Lunch) Status

To disentangle the correlated variables of ethnicity and income, we separate subsidized students from our sample, and measure the achievement gap between minority and majority students *not* receiving the lunch subsidy. By excluding low-income students from the analysis, we eliminate some of the variability in income across our sample, which would presumably diminish the effects of income on the remaining outcomes.

When we contrast the results of the unsubsidized group with the full sample (Figure 5), we find that the median score for minorities rises in the unsubsidized group, an indication that low-income minorities do not perform quite as well as the minority in the unsubsidized groups. The statistics for the majority group are largely unchanged, which is largely due to the low number of majority students receiving the lunch subsidy.

	All Students	Unsubsidized
No. of Students Tested	1098	963
Minority Students	320	206
Minority %	29%	21%
Median (Min)	262	269
Median (Maj)	299	300
Rank Sum Index	-0.58	-0.50
Median Gap (Recal.)	-0.20	-0.18
Pct. Meet/Exceeds (Min)	72%	79%
Pct. Meet/Exceeds (Maj)	95%	95%

Figure 5: Achievement Gap Statistics excl. Low-Income - Cohorts of 2004-06

Excluding low-income students from the analysis does improve minority outcomes; however, it is clear that the magnitude of the racial gap among unsubsidized students is comparable to the overall gap. In other words, even when we remove low-income students from the analysis, most of the race-based gap remains.

6.2.3. Census Income Data

Despite the findings of the previous section, we seek more evidence regarding the relationship of income to the racial achievement gap. The lunch subsidy information allows us to identify students who live in a “low-income” household, as defined by the state. However, the indicator does not provide much information about students who do not qualify for the subsidy. If there is substantial variation in economic status among the unsubsidized students, then these differences may still contribute to the achievement gap.

The Oak Park community is notable for its ethnic diversity, and Census data indicate that that the community has a substantial amount of economic diversity as well. The lowest level of geography that the Census Bureau provides sample data for is the *block-group*; as

of the 2000 Census, there are 51 block groups in Oak Park⁴². Although the median family income in 2000 was \$81,703, median family incomes varied considerably across block-groups, ranging from a low of \$44,514 to a high of \$170,073.

6.2.3.1. Methods

Given the range of economic diversity in Oak Park, we use the Census data to estimate differences in achievement amongst students living in economically comparable areas of Oak Park. These analyses allow us to evaluate the hypothesis that differences in economic status are influencing the Oak Park minority gap; for example, if the achievement gap were notably smaller amongst the affluent, then we would have support for that hypothesis.

In our approach, we divide the data into deciles of economically comparable students. As mentioned before, Oak Park is comprised of 51 Census block-groups; we create “economically comparable” deciles by bundling the block-groups according to median family income (Figure 6). For example, Decile 1 consists of the students who live in the 5 block groups with the lowest median family income; Decile 2 is composed of the next 5 lowest block groups; and so on. Every decile is representative of 5 block-groups, with the exception of Decile 10, which contains 6 block groups.

⁴² To provide perspective on the size of a block-group, Oak Park covers an area of 4.7 square miles, and has a population of roughly 50,000. (Source: Census 2000)

Decile	Avg. Income
1	\$ 53,525
2	\$ 61,044
3	\$ 64,549
4	\$ 71,504
5	\$ 76,195
6	\$ 80,783
7	\$ 90,300
8	\$ 98,251
9	\$ 105,283
10	\$ 151,964

Figure 6: Family Income Deciles for Oak Park, 2000

6.2.3.1. Findings

Once we have defined our deciles, we calculate and compare the achievement gap among the students in each decile. We compute the gap metrics from the 8th grade ISAT exam. As a caveat to these findings, this analysis requires a mapping of each Oak Park student to his or her Census block-group, determined by the student's home address. Due to the availability of data, the data herein refer solely from the students in the 2005 cohort.

Our findings, summarized in Figure 7, provide *no indication* that the achievement gap gets smaller as economic status improves. The figure lists student counts, the rank-sum index, and the recalibrated median gap for every decile containing at least 5 minority students (i.e., one relative and one absolute gap measure). For context, the final column presents the data for the full 2005 cohort.

Decile	1	2	3	4	5	6	8	10	All
No. of Students	29	30	28	22	34	29	44	78	369
Minority Students	13	12	12	9	17	6	17	10	105
Minority %	45%	40%	43%	41%	50%	21%	39%	13%	28%
Rank Sum Index	-0.69	-0.64	-0.58	-0.68	-0.70	-0.93	-0.40	-0.58	-0.62
Median Gap	-0.17	-0.16	-0.15	-0.22	-0.43	-0.28	-0.16	-0.28	-0.22

Figure 7: ISAT Math Gaps by Decile, Grade 8 (2005 Cohort)⁴³

From left to right, median family incomes increase; if we assume that family incomes for the students within each decile are also increasing, there does not appear to be a strong evidence of a smaller gap among affluent students. Relative measures, such as the rank-sum index, fluctuate across the deciles, but there is no evidence of a consistent increase or decrease in the magnitude of the gap. In contrast to the original hypothesis, in absolute terms, it appears that the absolute gap might actually *increase* with income. In the three lowest income deciles, the absolute gap is at its smallest. Conversely, the larger absolute gaps occur as median incomes increase. In short, the fact that Blacks are proportionately less wealthy than Whites does little to explain the achievement gap: even when we make matched-comparisons within groups with nearly the same income, the gap persists.

To summarize our analyses regarding income and achievement, despite clear evidence of an achievement gap among low income students in Oak Park, and a strong degree of overlap in the minority and low-income communities, it appears that income differences do not sufficiently explain the racial achievement gap. When we compare the performance of subsidized and unsubsidized students in the district, we see that the poverty gap is as large, if not larger than the racial achievement gap. However, we find that the racial achievement gap among unsubsidized students is nearly as large as the overall gap. Further analysis of economic status using Census data indicates large variations in income within the Oak

⁴³ Deciles 7 and 9 contain fewer than 5 minority students, so we omit them from the table.

Park community; however, the race gaps remain at all levels of economic stratification. Next, we examine the role of gender in the Oak Park race gap.

6.3 GENDER

6.3.1. Overview

During our exploration of the achievement gap in Oak Park, discerning the achievement trends of minority boys (specifically, African American boys) was an issue of particular concern to the district. In particular, there was suspicion that African-American boys might account for a disproportionate share of the achievement gap. This concern coincides with a large body of evidence that African-American boys have more difficulty succeeding in the classroom than their female counterparts (Noguera, 2002). On the national front, the concern about the plight of minority boys is such that educators in several districts have taken the additional step of creating achievement programs specifically for minority boys (Hu, 2007).

As we do throughout this study, we use the district's achievement data to investigate the issue of interest, which is now the role of minority males on the overall Oak Park achievement gap. To determine whether minority males account for a disproportionate share of the gap, we compare the outcomes of minority males with the outcomes of minority females. However, to provide context for our findings, we also compare the achievement trends across gender among the majority group.

Our analysis of gender and achievement in Oak Park begins with a comparison of male and female achievement across ethnic groups. We will assess differences across gender using the same tools we have used to study the minority gap. First, we calculate absolute and relative measures of the difference in male and female performance. As with our study of the racial gap, the metrics for measuring gender gap are indexed on a range of -1 to 1; here, negative values will imply that males are at a disadvantage, and positive values will imply

that females are at a disadvantage. Again, a value of zero represents parity with respect to gender.

For ease of exposition, we limit our presentation to a representative subset of our measures. The rank-sum index provides a sense of the gap in relative terms. We measure the absolute difference in achievement is measured by comparing median scores for the group. Also, we include the “Meets/Exceeds” metric as an alternative measure of absolute performance and a more familiar approach to framing achievement.

Also, as we have done in **Chapter 5**, we compare changes in performance as they age among boys and girls of comparable skill. To review the approach, we group students according to their third-grade performance level (i.e., “Below Standards”, “Meets Standards”, or “Exceeds Standards”). Given their earlier performance, a student’s performance level on the eighth-grade exam will either be higher, lower, or the same as before. We compare the empirical transition rates for boys and girls and we test the significance of any variations across the genders. As before, we use the terms *Below3*, *Meets3*, and *Exceeds3* to classify students according to their Grade 3 performance level.

After we have developed a general sense of how the genders differ, we study gender and achievement trends *within* the minority and majority groups. We compute gap metrics and transition probabilities as before, noting any differences across the groups. As in **Chapter 5**, we draw these observations from the outcomes of Oak Park students who completed Grade 8 in 2004, 2005, or 2006 and were in the school district for a minimum of five years.

6.3.2 Gender and Achievement across Ethnic Groups

In **Chapter 5**, we found student performance to vary significantly with respect to ethnicity across several years of data; now, we use the same data to determine the extent of variation with respect to gender. We begin by calculate the absolute and relative gap metrics for

boys and girls, regardless of race. The measures were computed for all three ISAT exams (Grades 3, 5, and 8), and the findings are summarized in Figure 8.

The gap metrics indicate that the differences in male and female performance are quite small, race notwithstanding. The district has a nearly even split between boys and girls (53% vs. 47%) and the gap measures are all extremely small. There is an indication that boys might perform slightly better on the third grade exam, but by the time the students graduate and leave the district, most of the metrics indicate no difference in the performance of eighth grade boys and girls.

Notably, the Meets/Exceeds metric runs a bit counter to the other interpretations. The percentages for males and females are nearly equal in third grade, and afterward, the two groups drift apart, indicating that the gap grows rather than shrinks with time. It is likely that the conflict is likely due to the lack of precision in the Meets/Exceeds metric (as outlined in Chapter 3), coupled with the very small magnitude of the gender “gap”.

Comparison of Performance Metrics			
	c04-06 Math3	c04-06 Math5	c04-06 Math8
No. of Students	1150	1150	1150
Male Students	615	615	615
Male %	53%	53%	53%
Median (Male)	224	248	288
Median (Female)	220	248	288
Rank Sum Index	0.07	0.02	0.00
Median Gap (Recal.)	0.02	0.00	0.00
Pct. Meet/Exceeds (Male)	89%	86%	87%
Pct. Meet/Exceeds (Female)	88%	87%	90%

Figure 8: Math ISAT Performance by Gender (2004-2006 Cohorts)

Aside from the gap metrics, we also consider changes in progress among students of comparable skill. As we saw in the previous chapter, comparisons of overall group behavior can conceal significant differences among students of comparable skill. Figure 9 summarizes the changes in performance from Grades 3 to 8, noting the proportion of students whose performance level improved, declined, or stayed the same over time. Within each category of 3rd Grade performance, we use the Chi-Square test to test the significance of any discrepancy across genders.

District Progress - Mathematics				
Grade 3 Performance	Grade 8 Performance	Female	Male	p
All Students	Higher	17%	12%	0.00
	Same	71%	70%	
	Lower	12%	18%	
Exceeds Standards	Same	83%	78%	0.19
	Lower	17%	22%	
Meets Standards	Higher	29%	24%	0.03
	Same	63%	60%	
	Lower	8%	16%	
Below Standards	Higher	43%	36%	0.43
	Same	57%	64%	

Figure 9: Transition Frequencies by Gender (2004-2006 Cohorts)

As we have seen before, 3rd Grade performance is a strong predictor of 8th Grade performance; 70 percent of the time, students had the same performance level on both exams, regardless of gender. However, when student performance levels change, girls are more likely to improve (17 percent vs. 12 percent), whereas boys are more likely to decline (12 percent vs. 18 percent). Moreover, Chi-Square testing indicates that the difference is statistically significant, as indicated by a p-value of zero under the null hypothesis. The indication that girls have higher rates of progress over time coincides with the notion of girls catching up to boys over time, a trend we saw in the gap metrics in Figure 8.

With each performance level, the data show that females were more likely than males to either improve over time or to maintain high levels of performance. The students in the *Meets3* subgroup most influence the pattern in overall performance between males and females. In **Chapter 5**, the *Meets3* subgroup also exhibit the most influence over overall differences in achievement between minority and majority students.

6.3.3 Gender and Achievement within Ethnic Groups

Although there is some evidence that males hold a slight edge in performance in the third grade, females appear to erase any hint of a gender gap by eighth grade. In closing the gap, it appears that females make more progress from Grade 3 to Grade 8 than their male counterparts. These findings certainly allow for the possibility that all males (including minority males) may be falling behind in the district. Here, we compare the performance of boys and girls within our majority and minority ethnic groups.

6.3.3.1. Gender and Achievement among the Majority Group

When we calculate achievement gap metrics among boys and girls in the majority group (Figure 10), we find evidence of a small gender gap favoring the boys. Of the various metrics available, relative metrics (such as the rank-sum index) provide the strongest indication of any difference in performance between males and females. In relative terms, the boys are slightly ahead of girls on the third grade exam, and the advantage wanes by the 8th grade exam. The absolute metrics, on the other hand, give no indication of a gap, which implies that although the boys are technically ahead of the girls from grades 3 to 8, there was never any real “distance” between male and female scores.

Comparison of Performance Metrics (Majority Group)			
	c04-06 Math3	c04-06 Math5	c04-06 Math8
No. of Students	778	778	778
Male Students	429	429	429
Male %	55%	55%	55%
Median (Male)	231	259	300
Median (Female)	228	255	298
Rank Sum Index	0.12	0.08	0.03
Median Gap (Recal.)	0.02	0.01	0.02
Pct. Meet/Exceeds (Male)	96%	95%	95%
Pct. Meet/Exceeds (Female)	95%	95%	95%

Figure 10: Performance by Gender (MAJ Group), 2004-2006 Cohorts

We list the transition rates for the majority group in Figure 11. Within subgroups, the differences in gender indicate that in all cases, girls in the majority group make at least as much, if not more, progress as their male counterparts. Although the differences within individual skill groups were not large enough to refute the hypothesis of no gender effect on transition rates, the difference in overall progress is statistically significant.

District Progress				
Grade 3 Performance	Grade 8 Performance	MAJ Female	MAJ Male	<i>p</i>
All Students	Higher	15%	11%	0.02
	Same	75%	74%	
	Lower	10%	16%	
Exceeds Standards	Same	86%	81%	0.13
	Lower	14%	20%	
Meets Standards	Higher	35%	31%	0.33
	Same	60%	61%	
	Lower	5%	9%	
Below Standards	Higher	47%	47%	1.00
	Same	53%	53%	

Figure 11: Performance Level Transition Frequencies by Gender (Majority Group)

6.3.3.2 Gender and Achievement among the Minority Group

In contrast to the majority group, the gap metrics for minorities (Figure 12) indicate that gender differences in achievement grow over time. In the third grade, minority girls slightly outperform the boys, and the advantage appears to grow as the students get older. The Rank-Sum Index, indicates that, in relative terms, the populations move apart from one another. However, the absolute measures indicate that the distance between the populations remains fairly small. Nevertheless, girls outperform boys in the minority group on every exam, and regardless of the metric, the gap is larger in eighth grade than it is in third grade. Thus, minority males consistently exhibit the lowest levels of performance on the ISAT exam.

Comparison of Performance Metrics (Minority Group)			
	c04-06 Math3	c04-06 Math5	c04-06 Math8
No. of Students	320	320	320
Male Students	165	165	165
Male %	52%	52%	52%
Median (Male)	196	224	256
Median (Female)	200	226	266
Rank Sum Index	-0.04	-0.13	-0.15
Median Gap (Recal.)	-0.02	-0.01	-0.04
Pct. Meet/Exceeds (Male)	70%	63%	66%
Pct. Meet/Exceeds (Female)	73%	71%	77%

Figure 12: Performance by Gender (MIN Group), 2004-2006 Cohorts

A comparison of minority transition rates (Figure 13) reveals that similarly situated minority girls make more progress than the boys. As we saw in the majority group, the largest differences in the behavior of minority girls and boys over time are within the *Meets3* group. Nearly seventy percent of females in the *Meets3* group continue meeting standards in Grade 8, and among students who move, over half of the females improve

rather than decline. In contrast, sixty percent of minority males in the *Meets3* group remained at that level in Grade 8, indicating that the males were more likely than females to change performance levels. Also, among the 41 percent of minority males who move away from the “Meets Standards” group, nearly 2 out of 3 fall below standards on the 8th grade exam.

District Progress				
Grade 3 Performance	Grade 8 Performance	MIN Female	MIN Male	<i>p</i>
All Students	Higher	21%	17%	0.18
	Same	64%	61%	
	Lower	15%	22%	
Exceeds Standards	Same	63%	54%	0.49
	Lower	38%	46%	
Meets Standards	Higher	19%	14%	0.09
	Same	68%	60%	
	Lower	14%	27%	
Below Standards	Higher	42%	33%	0.36
	Same	58%	67%	

Figure 13: Performance Level Transition Frequencies by Gender (Minority Group)

Despite these discrepancies among the boys and girls in the minority group, the differences were not statistically significant, according to our testing method. Recall that we evaluate the differences in transition rates using the Chi-Square test; for each of the four groups in the table, the value *p* represents the likelihood of the observed outcome, under the null hypothesis that males and females behave identically. Generally, we reject the null hypothesis when *p* is less than 0.05; in this case, the *p* –values in Figure 13 are too large to rule out the possibility that the gender differences in our empirical data are the result of measurement error.

6.4. CHAPTER SUMMARY

Within the middle-class community of Oak Park, we have encountered two issues with the potential to contribute to the achievement gap. Student mobility and poverty, known correlates of student achievement, have a disproportionate presence among underrepresented minority students. When compared to White and Asian students, minority 8th graders are also more than twice as likely to have transferred to the district within the last five years. Also, minorities are 9 times more likely to have come from “low-income” households, as defined by the school district.

However, these socioeconomic disadvantages do not come close to explaining the Oak Park gap. Despite the correlation between low income and low achievement, academic performance among minorities does not generally improve with income, and thus, the achievement gap exists across all income groups. Regarding student mobility, a comparison of veteran and recently enrolled minority students suggests that student mobility plays a relatively minor role in explaining the Oak Park gap. Students with five years or more within the district outperform newcomers, regardless of the student’s race. When we measure the gap across all Oak Park students, the data show that the magnitude of the overall gap is very close in size to the magnitude of the gap among long-term residents.

We also examined the role of gender in shaping achievement patterns in the district. Although there are statistically significant differences in the academic outcomes of boys and girls, the “gender gap”, within either ethnic group, is fairly small. Regardless of ethnicity, girls appear to make larger achievement gains than the boys between 3rd and 8th grade. In the case of the majority group, boys outperform girls early, and the girls “catch up” academically by the time they graduate from the district. Within the minority group, girls outperform boys in the third grade, albeit by a small margin; the minority girls also exhibit higher levels of performance over time, causing minority boys seem to fall even further behind the girls.

In summary, these findings indicate that these correlates of low achievement do not adequately explain the achievement gap in Oak Park. However, we recognize that even if there were evidence that student mobility or income gaps were driving the achievement gap, the school district would have little power to affect the trend. For example, we can not expect the school district to lower the mobility rate of minority students, or increase the wealth of minority families.

In **Chapter 7**, we examine the district's ability to reduce achievement gaps. Regardless of external factors like mobility and income, there are several aspects of a student's education that the district can affect. We evaluate several reform strategies based on the potential for underlying data to predict the Oak Park minority gap. Also, we study the role of teachers in predicting student performance.

7. Potential Levers for Reducing Achievement Gaps

Through various analyses, we have been able to identify some trends regarding the achievement gap faced by the district. In this chapter, we perform other analyses designed to identify areas of opportunity and inform the district's efforts to narrow the gap. The analyses in this chapter represent a data-driven response to several questions raised by the district concerning student achievement, stated below:

- *How do different transition rates across performance levels predict the 8th grade gap?*
- *How do gaps in Reading comprehension predict the Mathematics gap?*
- *How does the existence of honors Math classes predict the gap?*
- *Do student gains in achievement vary significantly by teacher?*

Where appropriate, we use empirical data to quantify how these factors predict student achievement over time. While we have used multiple gap measures throughout the thesis, we simplify the discussion by using a single metric, the difference in average score. Our baseline is the historical eighth grade gap on the ISAT Math Exam.

7.1. ACADEMIC TRANSITION RATES

7.1.1. Overview

In Chapter 5, we saw significant differences in the progress made during elementary school between minority and majority students. In the end of Chapter 6, we compared progress

rates across gender within racial groups and, for both groups, girls made more progress than boys. Together, if the Oak Park school leadership were to focus its efforts on improving the performance of minority males, how might these improvements predict change in the overall achievement gap?

The data indicate that girls in Oak Park tend to make more progress than boys, regardless of gender. In the combined data, and within the majority group, we saw evidence of a small gender gap, favoring boys, at the third grade, but which disappeared by the eighth grade. In contrast, the minority gap metrics show a small gender gap, favoring girls, which grows as the students get older.

Comparison of the transition rates of minority males to the rates for other groups (Figure) provides further confirmation that minority males make less progress than minority females, majority males, and majority females. Among students who perform below standards in the third grade, minority males are the most likely to continue performing below standards in the eighth grade. Of the students that meet the Grade 3 standards, minority males are the most prone to decline and perform below Grade 8 standards. Also, among students exceeding the Grade 3 standards, minority males are most likely to drop to a lower performance level in Grade 8 standards.

District Progress - Mathematics					
Grade 3 Performance	Grade 8 Performance	MAJ Female	MAJ Male	MIN Female	MIN Male
Exceeds3	Same	86%	81%	63%	54%
	Lower	14%	20%	38%	46%
Meets3	Higher	35%	31%	19%	14%
	Same	60%	61%	68%	60%
	Lower	5%	9%	14%	27%
Below3	Higher	47%	47%	42%	33%
	Same	53%	53%	58%	67%

Figure 1: Transition Frequencies by Ethnic Group, Gender (2004-2006 Cohorts)

Although it is clear that minority males had the least favorable outcomes, the most striking aspect of Figure is that the differences between the behavior of minority and majority students appear far greater than any gender based differences. We mention this point not to understate the plight of minority males, but to emphasize the point that, according to the data, the opportunity for improvement among minority females is nearly as large as the opportunity presented by minority males.

Now, suppose that the transition rates for minorities had been closer to those in the majority group. This scenario would have some implications, namely:

- Minority students with high third grade scores would be more likely to exceed the eighth grade standards.
- Minority students meeting the third grade standards would be less likely to fall below standards in eighth grade, and more likely to exceed the eighth grade standards

Any of these consequences would imply that eighth grade test performance had improved for some minority students, and, holding scores constant for non-minorities, there is a likelihood that the eighth grade achievement gap would decrease. However, if the district were able to narrow the eighth grade gap this way, how much improvement would we see in the data?

7.1.2. Methods

In addition to providing context to the performance trends in Oak Park, the transition frequencies arguably serve as a measure of district effectiveness. Although many factors can account for the change in a student's academic performance over time, the district assumes a leadership role in guiding a student's development, particularly by the third grade. Thus, the transition frequencies reflect the district's ability to promote performance gains among its students.

Considering transition rates as a measure of effectiveness is essential to the notion that, through appropriate interventions, the district could become more effective with its minority students, a move that would improve eighth grade minority test scores and potentially reduce the eighth grade achievement gap. For a group of students, we define the relationship between their transition rates (f) and their mean Grade 8 performance (X) with the following equation:

$$X = \sum_{i,j} p_i f_{ij} X_{ij}$$

The equation presents mean achievement for a group of students as a weighted average of the eighth grade means for each transition; in Figure 2, we show the components of the equation using the data from minority males. In this formulation, i denotes the Grade 3 performance level (Below, Meets, or Exceeds), and j denotes the relative performance on the eighth Grade exam (Higher, Lower, Same). The p_i term represents the percentage of students who were at performance level i on the third grade exam. For a given performance level, the f_{ij} and X_{ij} terms, respectively, represent the likelihood of making transition j , and the mean eighth grade score attained by the students who made that transition.

Transition Data and Mean 8th Grade Scores, MIN Males				
3rd Grade Performance <i>(i)</i>	Distribution <i>p(i)</i>	Transition Type <i>(j)</i>	Rate <i>f(i,j)</i>	Mean ISAT Score <i>X(i,j)</i>
Exceeds Standards	0.17	Same	54%	313.7
		Lower	46%	275.2
		Higher	14%	303.7
Meets Standards	0.53	Same	60%	266.5
		Lower	26%	234.9
Below Standards	0.30	Higher	33%	252.6
		Same	67%	231.0
Group Mean (X):				261.3

Figure 2: Transition Data and Mean ISAT8 Performance, MIN Males

Using the historical transition rates and performance data for Minority boys, Minority girls, Majority boys, and Majority girls, we adjust the transition rates to estimate how differences in the rates across race and gender predict the eighth grade achievement gap. Of the four groups, Majority females had the best transition rates; as such, we assume these rates to be “best-in-class” for the district and set them as an upper limit for our adjustments.

As we make these adjustments, it is important to note that even when transition rates are equal for two groups, there is still the possibility of an eighth grade achievement gap. When achievement gaps appear early, as they do in Oak Park, establishing equal transition rates for the future is unlikely to close the gap. For this analysis, we assume that values of p_i are fixed by race and gender; accordingly, we do not expect equal transitions going forward to make up for initial gaps in achievement.

7.1.3. Findings

We summarize the findings in the context of three “what-if” scenarios; for each scenario, we note the adjustments made to transition frequencies in the Minority group, and estimate their ability to predict change in the eighth grade gap. In this context, the achievement gap equals the weighted average of the eighth grade scores for Majority males and females, minus the weighted average of the eighth grade scores for Minority males and females. Unless otherwise stated, we assume throughout the scenarios that the transition rates for Majority students (boys and girls) remain constant.

We discuss the following three scenarios:

1. Minority males attain the exact performance levels in eighth as they had in third (i.e., no movement a higher or lower level over time.)
2. The transition rates for Minority males are set equal to the transition rates for Minority females.

- The transition rates for Minority males are set equal to the transition rates for Majority males, and the transition rates for Minority females are set equal to the transition rates for Majority females.

We create multiple scenarios by adjusting the transition rates of minority males, minority females, or both; the remaining groups retain their historical transition rates (c.f. Figure).

For each scenario, we describe predictive power in terms of the estimated reduction to the existing eighth grade gap. To illustrate, suppose that Majority group average on the eighth grade exam was 300, and the Minority group average was 260, a nominal difference of 40 points. If transition rates for the Minority group were adjusted such that the overall average score for Minority increased by 4 points, then the eighth grade achievement gap would be $4/40 = 10\%$ smaller.

As summarized in Figure 3, the data indicate that each of the proposed scenarios would reduce the gap.

Summary of Transition Effects on 8th Grade Testing Gap					
	Mean 8th Grade Score				Reduction in Gap
	MIN Male	MIN Female	MIN - All	MAJ - All	
Actual	261.3	268.2	264.7	297.6	-
Scenario 1	263.9	268.2	266.0	297.6	4%
Scenario 2	265.6	268.2	266.9	297.6	7%
Scenario 3	270.3	274.9	272.5	297.6	24%
Sample Size	165	155	320	778	

Figure 3: Potential Reduction in Gap from Adjusting Transition Frequencies

Scenario 1: (Grade 3 distribution = Grade 8 distribution for MIN males)

For minority males, the Grade 3 distribution of performance is actually more favorable than the Grade 8 distribution of performance. As shown in **Chapter 6.4**, the percentage of minority males meeting or exceeding the state standards was 70 percent in the third grade; by the eighth grade, the proportion fell to 66 percent. In the first scenario, we adjust the Grade 8 distribution to match the Grade 3 distribution and estimate related changes in the achievement gap. In other words, we will assume that each minority male attained the same performance level in Grade 8 as he had in Grade 3.

Under this assumption, had minority males maintained the same levels of performance on the eighth exam as were established on the third grade exam, the mean difference in score on the eighth grade exam would have been reduced by 4 percent. Although the performance of minority males does decline after third grade, the effect of this decline on the eighth grade gap appears marginal.

Scenario 2: MIN Males transition rates equal MIN female rates

In the second scenario, we set the transition rates for minority males equal to the transition rates for minority females. The transition rates for the majority groups are unchanged. Under this scenario, the overall race gap would decrease by about seven percent. The modest reduction in the race gap reflects the fact that, in this district, differences in performance among minority males and females are relatively small compared to the difference in majority and minority performance. This finding indicates that interventions strictly focused on minority males play a limited role in predicting race gaps in Oak Park.

Scenario 3: MIN transition rates equal MAJ transition rates

Lastly, we consider a scenario in which minority males and females had made equal progress with the majority group. We allow differences in transition frequencies with respect to gender, but not race; in other words, we assume that minority males have the transition rates of the majority males, and we assume that minority females have the transition rates of majority group females.

Under this scenario, we reduce the nominal Grade 8 gap by 24 percent. This figure indicates that, had Oak Park been able to maintain progress among minorities as well as it had with the majority group, the eighth grade gap would have been about three-fourths of its measured size. This finding supports the indication in **Chapter 5** that the achievement gap is present by the third grade, and grows over time in Oak Park.

To the extent that the school district can influence the academic transitions that students make, these results also indicate that if the district could move the transition rates of minority third-graders closer to the transition rates for the majority group, there would be a reduction in the eighth grade testing gap. The data indicate that, had there been no difference in the transition rates, Oak Park would have had a higher proportion of minority students meeting or exceeding standards on the eighth grade exam. Under the scenario of no difference in transition rates across ethnic groups, we estimate a reduction in the existing eighth grade testing gap of nearly 25%. We note that this estimate takes into account the fact that the Minority group is already lagging behind by Grade 3, and that the improvement in this case would come primarily from more effective interactions with Minority students who have already met or exceeded expectations on the third grade exam. Even still, the large gap in the third grade data makes it clear that improving transition rates beyond the third grade will not *close* the achievement gap in Oak Park.

Despite the widespread concern regarding the academic welfare of Minority males, the data indicate that any gender related gaps in the district are secondary to the ethnic gap. However, it would be a mistake to infer that Minority males in Oak Park would not require specialized intervention. To be sure, identifying and addressing the particular needs of Minority males will be an important strategy in closing the achievement gap. However, we point out that if Minority males are at a higher level of academic risk than Minority females, the incremental effect of that risk on the testing gap is rather small. Thus, we anticipate that *interventions that address the needs of minority boys and girls will be far more effective in closing the race gap than efforts that focus solely on minority boys.*

7.2 READING COMPREHENSION

7.2.1. Overview

In this section, we consider the possible connection between Reading skills and the observed achievement gap in Mathematics. We mention the Reading gap here because its presence may contribute to the Mathematics gap. There is a hypothesis among members of the school district that a causal relationship exists between Reading proficiency and student performance on the Mathematics exam. If this were true, then the lower Reading proficiency observed among minorities (i.e., the Reading gap) might constitute a barrier to progress in Mathematics over time. In **Chapter 5**, we saw evidence that minority students tended to make less progress than their counterparts in the majority group; this hypothesis raises the possibility that lower Reading skills may help explain the trend.

In the context of written exams, the potential for Reading skills to predict general test performance has an intuitive appeal. A student with poor Reading comprehension would almost certainly be at a disadvantage when taking any type of written exam, regardless of the test subject. In a more general sense, some researchers suggest that a child's propensity to read predicts improved academic performance. To this point, a recent study by the National Endowment for the Arts notes a positive correlation between a student's access to reading materials (as measured by the number of books in the home) and performance on the Mathematics portion of the National Assessment of Educational Progress (NAEP).⁴⁴

The best available proxy of Reading skills in Oak Park are the data from the Reading Comprehension component of the ISAT state exam. Like the Mathematics exam, all of the students in our sample took the state Reading exam in Grades 3, 5, and 8. We compare trends in the two subjects with summary data in Figure 4.

⁴⁴ National Endowment for the Arts, *To Read or Not to Read* (2007)

In short, the data show that Oak Park also has an achievement gap with respect to the Reading exam, albeit with slightly different characteristics. Although scale scores for the Reading and Mathematics appear to be similar in magnitude, we cannot directly compare scores from different subject areas, because the scales have different calibrations. However, we can compare performance across subjects by using the gap metrics identified earlier in this study.

	c04-06 Math3	c04-06 Math5	c04-06 Math8	c04-06 Read3	c04-06 Read5	c04-06 Read8
Median (Min)	199	224	262	196	222	243
Median (Maj)	231	257	299	229	255	269
Rank Sum Index	-0.58	-0.59	-0.58	-0.57	-0.55	-0.56
Median Gap (Recal.)	-0.15	-0.09	-0.20	-0.15	-0.17	-0.08
Pct. Meet/Exceeds (Min)	72%	67%	72%	59%	62%	69%
Pct. Meet/Exceeds (Maj)	96%	95%	95%	91%	92%	95%

Figure 4: Comparison of Testing Gaps in Mathematics and Reading

For example, in the third grade, the mathematics gap and the reading gap are basically identical in absolute and relative magnitude. As students move from Grade 3 to Grade 8; relative gaps, such as the Rank Sum Index, remain about the same on both exams. However, while absolute gaps, like the median gap, grow over time in Mathematics, the absolute gaps for Reading appear to get smaller. The Meets/Exceeds data shows a gradual increase in the percentage of minority students meeting or exceeding the state standards; however, those same students appear to more or less hold their ground with respect to the Math standards.

7.2.2. Methods

Using district data, we investigate whether the data at hand are consistent with a hypothesis that deficits in Reading comprehension inhibit gains in Mathematics performance. For this analysis, we measure Math progress as the difference in the Grade 8 and Grade 3 Math exam scores. Our proxy for Reading comprehension will be the reading score from the

Grade 5 exam. We chose to use the 5th grade Reading score because we wanted to a measure of Reading skills that was gathered about halfway between the initial (third grade) and final (eighth grade) state assessments of Math performance.

If reading skills correlate with achievement in Mathematics, then we would expect that, with all other factors equal, students with high reading scores in 5th Grade would make greater progress in Math between the third and eighth grade than would students with lower Grade 5 reading scores. Such a correlation would not prove that improved reading scores lead to larger math gains, but it would increase the plausibility of arguments that claim a causal link.

We test this hypothesis using a single variable linear regression. Our regression model takes the following form:

$$M_{38} = \alpha(\text{Read5}) + \beta + \varepsilon$$

In this model, M_{38} represents improvement in math (that is, the difference in the Grade 3 and Grade 8 Math score), and Read5 represents the Grade 5 reading score. The symbols β and ε represent the intercept and error terms, respectively.

However, we constrain the analysis in several ways so that we are comparing M_{38} values for individuals who are similar except in reading proficiency. We have already noted statistically significant differences in Math progress by both ethnicity and gender (c.f., **Chapter 6.3**); to control for these factors, we perform *separate regression analyses* for minority males, minority females, majority males, and majority females.

In our approach, we identify binary correlates of achievement (i.e., race and gender), and partition the data along these dimensions *before* carrying out the regression. An alternative approach would control for ethnicity and gender by adding binary variables to the regression model. Rather than partitioning the data and performing multiple analyses, this

method would entail a single regression on the full dataset. Despite the fact that our approach yields smaller samples for regression, we chose this approach because the binary-variable approach entails additional assumptions that might not hold.

In addition to ethnicity and gender, we consider another factor that may predict math gains regardless of the reading exam: the Grade 3 math score. Suppose we had two students with identical reading scores on the Grade 5 exam. If one of the students had scored thirty points higher on their Grade 3 exam, this information could certainly influence our perception of who would make the larger gains over time. For example, the high scoring student has less room for incremental improvement, so the low-scoring student might be in a better position to make large gains.

We account for the third grade Math score in the analysis in the same way we account for race and gender; rather than including the score in the regression model, we only compare students with *highly similar scores* on the third grade Math exam. We create similar groups by sorting our data into deciles based on Grade 3 Math scores⁴⁵. Once we have done so, we perform the regression analysis within each decile. Because data within deciles are independent, this approach provides us with $(10 \text{ deciles}) \times (2 \text{ genders}) \times (2 \text{ ethnic groups}) = 40$ independent analyses of the relationship between reading comprehension and math gains.

When a dataset of approximately 1100 students is broken into 40 independent groups, each regression estimate, on average, comes from a group of fewer than thirty students. As a result, many of the analyses are subject to a non-trivial degree of measurement error. On its own, a regression analysis with such a high degree of error would provide very little evidence of a meaningful connection between reading scores and math progress. However, when we consider the *collective* findings from several regression analyses within a given race and gender, some stronger patterns begin to emerge.

⁴⁵ Other groupings are possible; for example, an alternative approach might group students within evenly-spaced score intervals.

7.2.3. Findings - Minority Males

We begin with the results for minority males, summarized in Figure 5. The columns, labeled one through ten, represent increasing deciles of prior (Grade 3) Math performance. The first three rows display the number of students per decile and the range of Grade 3 scores that define each decile. The next two rows indicate the coefficient estimate for the reading score and the associated t-value (that is, the coefficient, divided by standard error).

Taken individually, most of the regression data are inconclusive. Sample sizes range from 10 to 23, relatively small groups that increase the likelihood of measurement error. The coefficient varies considerably from one decile to another, and in many cases, the associated t-values are small in magnitude. In general, a t-value of ± 2 or greater indicates that a coefficient estimate has statistical significance⁴⁶; conversely, small t-values indicate the absence of statistical significance, meaning that the possibility of no correlation between reading score and the math gains cannot be excluded.

Decile	1	2	3	4	5	6	7	8	9	10	Sum
N	19	16	23	10	15	17	15	19	14	17	165
min	124	167	177	186	190	198	208	214	222	230	
max	165	175	185	188	196	206	212	220	228	276	
coefficient	0.27	0.07	-0.17	0.45	0.48	0.69	0.61	0.00	0.24	0.55	Average 0.32
t-value	1.51	0.65	-1.05	1.43	1.57	4.97	3.37	-0.01	1.06	3.35	

Figure 5: Regression Results by Decile – Minority Males, All Cohorts

The individual regression analyses are not convincing, but when we look at the findings in tandem, the evidence of correlation becomes stronger. If there is no relationship between reading scores and math progress, then the “true” coefficient of our regression model would equal zero. If this is true, then, given a large number of independent regressions, we

⁴⁶In regression analysis, the t-value assesses the likelihood that a nonzero coefficient is statistically significant. For samples larger than 30, a t-value of ± 2 is roughly equivalent to a 5% probability that the regression coefficient is the result of random error.

would expect the estimated coefficients to average to zero. Because the regression estimates for each decile are mutually independent, so we would also expect our coefficients estimates to be positive or negative with equal likelihood.

When we plot the slope coefficients for minority males (Figure 6), we see that this was not the case. The regression analysis yields a positive coefficient in eight of ten cases, whereas, in the case of no correlation, we would expect a more even mix of positive and negative results. Because the analyses are independent, this outcome is analogous to flipping ten coins, assumed to be fair, and observing eight “heads” (or eight “tails”, for that matter). In such an experiment, the probability of observing eight or more of either outcome is about 11 percent (i.e. the p-value of the result is 0.11). This indicates that although the outcome could occur with a fair coin, the likelihood is about 1 in 9, which is somewhat low. In a similar vein, our analysis indicates, even if there is no correlation between the reading score and the math gains, we have a 1 in 9 chance of seeing this many positive coefficient estimates. Although this is an uncommon scenario, the evidence is not strong enough to reject the possibility that reading skills and math gains are uncorrelated.

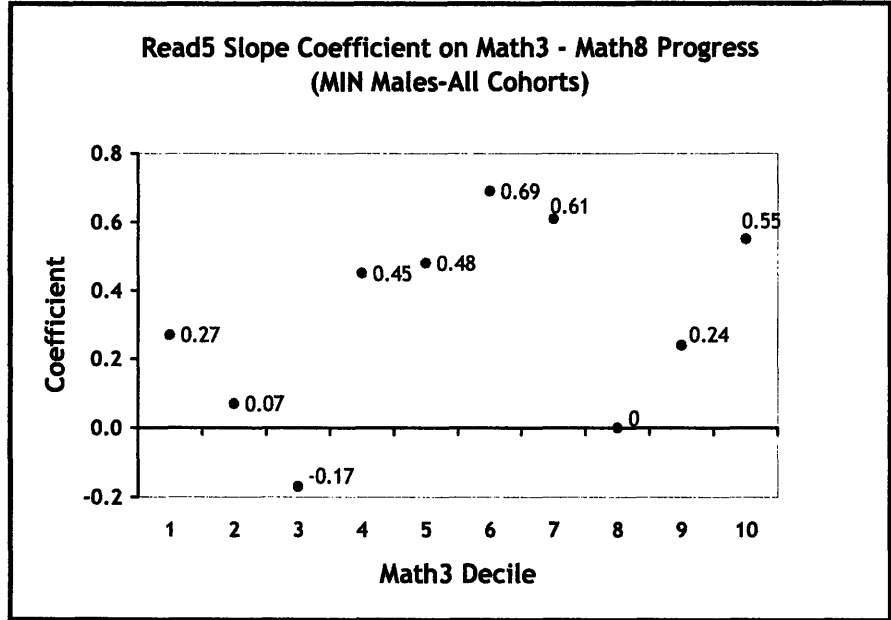


Figure 6: Plot of Regression Coefficients by Decile (Minority Males)

7.2.4. Findings for other groups

As we repeat the analysis for minority females and the majority groups, the findings resemble the results for minority males, providing further evidence of a real correlation between reading scores and math gains. The tables in Figure 7 summarize the regression results for minority females, majority males, and majority females.

In each of these groups, positive coefficients dominate negative coefficients when we compare across deciles. For minority females, eight of the ten deciles show a positive coefficient, indicating a positive correlation between reading scores and math gains. Among the majority group, *all ten* of the coefficients are positive for males, and nine of ten are positive for females. In total, thirty-five of forty regression analyses yielded positive coefficients.

To revisit our earlier analogy, observing this outcome under our testing hypothesis (i.e., that reading scores and math gains are uncorrelated) is similar to tossing a fair coin 40

times and observing 35 heads. Under the assumption that the Reading scores and Math gains are uncorrelated, the probability of observing a result this lopsided is less than 1 in 700,000. As a result, we have extremely strong reason to believe that, among students who are similar with respect to gender, ethnicity, and third grade Math score, Reading scores correlate positively with Math gains.

Regression of Math3 - Math8 Progress on Read5 Scores (MIN Group - Female)											
Decile	1	2	3	4	5	6	7	8	9	10	Sum
N	18	18	14	14	16	13	21	11	16	14	155
min	142	169	179	186	190	202	210	218	226	241	
max	167	177	185	188	200	208	216	224	237	276	
coefficient	0.41	0.08	-0.18	0.38	-0.12	0.49	0.25	0.08	0.02	0.33	Average 0.17

Regression of Math3 - Math8 Progress on Read5 Scores (MAJ Group - Male)											
Decile	1	2	3	4	5	6	7	8	9	10	Sum
N	43	51	39	46	38	60	33	40	38	41	429
min	153	196	214	222	228	233	243	249	257	267	
max	194	212	220	226	231	241	247	255	265	276	
coefficient	0.08	0.13	0.36	0.56	0.34	0.06	0.48	0.31	0.30	0.43	Average 0.30

Regression of Math3 - Math8 Progress on Read5 Scores (MAJ Group - Female)											
Decile	1	2	3	4	5	6	7	8	9	10	Sum
N	36	34	39	32	36	36	41	28	34	33	349
min	120	196	208	214	222	230	235	243	249	261	
max	194	206	212	220	228	233	241	247	259	276	
coefficient	0.20	0.15	0.24	0.26	-0.05	0.24	0.14	0.40	0.46	0.63	Average 0.27

Figure 7: Regressions by Decile – MIN Females, MAJ Males, MAJ Females - All Cohorts

But it is not enough to say that there is a positive correlation between reading proficiency and math gains. How much does a unit improvement in reading increase the predicted improvement in Math? For every ethnic and gender subgroup, the coefficient estimate varies across deciles; therefore, the best summary statistic for a given subgroup might be their arithmetic mean. For the majority group, the averages are quite close for males and

females (0.30 and 0.27, respectively). The average for minority males is 0.32, which is also quite close to the majority group. The average for minority females is noticeably lower than the others at 0.17; however, a difference of means test indicates that the deviation is not statistically significant.

Under the assumption that there is a causal relationship under which greater reading proficiency increases progress in Mathematics, these empirical findings imply that improved skill in reading does contribute to improvements in math over time, and subsequently, to higher math scores on the Grade 8 exam. Across the different subgroup, the average slope of math improvement when regressed against 5th grade reading is close to roughly 0.27. This statistic suggests that, if a student had scored ten points higher on their Grade 5 reading exam, the improvement in math between Grades 3 and 8 would grow by roughly three points.

7.2.5. Reading Scores and Math Achievement

When we extend this idea to the minority students in our sample, the reasoning suggests that if we were able to narrow the reading gap across ethnic groups, we could inhibit the growth of the math gap over time, resulting in a smaller gap in Grade 8 math performance. To estimate the predictive power of removing the reading gap, we will reference the average exam scores from our three cohorts of student data. For the majority group, the average score on the Grade 5 reading exam (*Read5*) was 254. In contrast, the *Read5* average for minority males was 218, a difference of 36 points. For minority females, the *Read5* average was 228, a difference of 26 points.

Now, suppose that both minority groups (male and female) had performed just as well as the majority group on the Grade 5 reading exam⁴⁷. In this scenario, the *Read5* average for minority students would also equal 254, thereby eliminating the Grade 5 reading gap. The

⁴⁷ This scenario that follows is illustrative of a situation in which there would be no Grade 5 reading gap. The condition that minority males and females perform equally well is not necessary, but it facilitates the discussion.

Read5 average for minority males would increase by 36 points; with an average coefficient of 0.32 for these students, this translates to an increase in the average *Math8* exam score of approximately 11.4 points. By similar logic, the average *Math8* score for minority females would increase by 4.5 points (26 point improvement \times 0.17 coefficient avg.). In total, the minority group average on the *Math8* exam would increase by about 8 points.

Under this scenario, we estimate the role of erasing the 5th grade Reading gap in predicting change in the 8th grade Math gap (Figure 8). In the figure, the final column of the table indicates the estimated improvement in eighth grade Math performance. In terms of average performance, the achievement gap drops from 33 points to 25 points, which corresponds to a reduction in the Grade 8 Math gap of nearly 25 percent, in terms of scale points. The absolute mean gap relates directly to the mean scale scores, so the magnitude of the absolute mean gap would also decrease by about 25 percent. The effect on the relative gaps is more difficult to determine, but it is reasonable to assume that the relative Math gaps will get smaller, but not disappear. Thus, under the causality assumption, we believe that eliminating the 5th grade Reading gap could reduce later gaps in Math considerably. However, we also note that a substantial Math gap would still remain.

Ethnic Group	N	Read5*	Coef.	E[inc]	Math8*	Potential Math8**
MIN Male	165	218	0.32	11.4	261	273
Female	155	228	0.17	4.5	268	273
MIN All	320	223	-	8.1	265	273
MAJ All	778	254			298	298

* Average performance, 2004 - 2006 Cohorts
 ** Estimated impact of no *Read5* gap on *Math8* score

Figure 8: Estimated changes to the *Math8* gap due to removal of the *Read5* gap

Although these findings are not conclusive, the analysis indicates that differences in reading skills during the intermediate years might account for up to one fourth of the

observed absolute difference in 8th grade Math performance. To be clear, our intent is not to suggest that reducing the Reading gap is any less challenging than reducing the Mathematics gap. However, the findings suggest that the successful early implementation of reading-based initiatives in the district may ultimately coincide with reduced achievement gaps in Mathematics.

7.3. MIDDLE SCHOOL HONORS ENROLLMENTS

7.3.1. Middle School Math in Oak Park

In Oak Park, students attend elementary school until the completion of fifth grade, at which point they transition to one of two middle schools.⁴⁸ Historically, the district has maintained a more detailed data set at the middle school level; in addition to test scores, there are course data for middle school students, such as teacher information and academic grades. We wish to incorporate these factors into our study of achievement gaps, but we lack sufficient data to do so at the elementary school level. As a result, the analyses in this section, and in the following section on teachers, address changes in the gap that occur during middle school. Accordingly, we constrain our analysis to students who have attended schools in Oak Park since the third grade and have complete middle school records⁴⁹.

Throughout elementary school, the district's core curriculum provides for a single Math course per grade level⁵⁰. However, during middle school, the district's core curriculum expands to two Math courses per grade level; students enroll in the different courses on the

⁴⁸ Prior to Fall 2002, students remained at the elementary school until the end of sixth grade, and attended middle school for Grades 7 and 8 only.

⁴⁹ Approximately 11% of the students in our test score dataset have incomplete or ambiguous middle school data. Of the 124 incomplete records, nearly 80% came from one school, a disproportionate amount given that the schools are about the same size. Unfortunately, this data loss is also disproportionate with ethnic groups, affecting fifteen percent of our minority records, in contrast to ten percent for the majority group.

⁵⁰ Every year, there are a number of students who do not follow the core curriculum due to academic, social, or perhaps disciplinary issues; to address their needs, the district provides alternative math courses. Over 97% of the students in our sample remained with the core curriculum throughout middle school.

basis of their prior Math performance. After students complete the regular sixth grade course (“Math 6”), the district offers a regular and an honors level math course for Grades 7 and 8. The regular course offerings for Grades 7 and 8 are “Math 702” and “Math 802”, respectively. By a similar token, the honors courses are “Math 703” and “Math 803.”

7.3.1.1. Enrollment Patterns by Ethnic Group

Intuitively, if a larger proportion of majority students take the honors courses in seventh and eighth grade, content and pedagogical differences between the regular and honors courses could exacerbate the Math gap in later years, particularly on the eighth grade state exam. Thus, it seems worth considering whether more minority students could plausibly enroll in honors Mathematics in middle school. We begin our examination of course enrollment by ethnicity by observing the enrollment patterns in our sample (Figure 9). In the figure, we depict the Oak Park middle school core Math curriculum as a network; the nodes are different courses, and the arcs represent student enrollment patterns from one year to the next. To show enrollment patterns, we label each arc with an ordered pair of probabilities; the first probability refers to the majority group, and the second probability refers to the minority group. From sixth grade, there are four distinct “paths” a student could “travel” through the core middle school curriculum.

As the data show, although majority and minority students take courses in the same schools, their paths through the Oak Park curriculum tend to vary significantly. From sixth grade, over 70 percent of the majority students in our sample enroll in the honors seventh grade Math course. However, less than 30 percent of minority sixth graders enrolled in the honors seventh grade course.

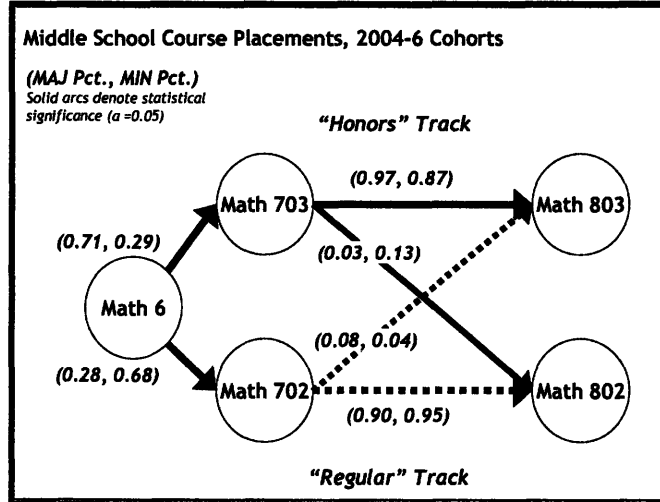


Figure 9: Middle School Course Placement, 2004-6 Cohorts

The seventh grade course placement is especially important because of the high correlation between the seventh grade placement and the eighth grade placement. This is true for both ethnic groups; however, it is worth noting that minorities in the honors seventh grade class were more likely than majority students to enroll in the regular eighth grade course (13 percent versus 3 percent). Under the assumption that the honors courses cover material in a greater depth, this pattern is another sign that high-performing minorities (i.e., those enrolled in the seventh grade honors class) are more likely to fall behind their majority peers over time (e.g., by enrolling in a less rigorous eighth grade course).

7.3.2. Course Enrollments and the Eighth Grade Achievement Gap

A comparison of course enrollments reveals that, after leaving the sixth grade, most of the Black students took regular math courses while in middle school, whereas most of the White students in the district were taking honors courses. If the black students were not able to cope with the honors classes, then putting them there in larger numbers might not reduce the achievement gap at all. But it is not certain that all black students who could enroll in honors Math are doing so.

In this section, we examine whether there is an opportunity to reduce the achievement gap we see in the eighth grade, when students leave the district. First, we examine the relationship between classroom performance and test performance in the eighth grade, while controlling for ethnicity and course enrollment (regular vs. honors enrollment). Next, we revisit the eighth grade classroom and testing gaps, this time, we focus on the differences in achievement among students enrolled in the same type of course.

7.3.2.1. Test Performance and Course Enrollment

We begin with an examination of the link between students' test scores in the eighth grade, and their classroom performances in the eighth grade. For classroom performance, we define a range of different levels of classroom performance, based on the student's Average Math Grade (AMG); that is, the average of the course grades received in their eighth grade Math course⁵¹.

In Figure 10, we have a table of classroom performance levels, mapped to different ranges of the AMG. The AMG ranges are not a precise replication of the Oak Park grading system; for example, the district would likely classify a student with a 3.99 AMG as an "A" student, rather than an "A-" student. However, the classroom level definitions allow us to segment the data in a manner that closely resembles the Oak Park approach.

⁵¹ Students in the district receive course grades every trimester, or three times a year. The Average Math Grade is the average of the trimester grades; see Section 4.4.2

Category	AMG
A	4.00 and above
A-	3.50 - 3.99
B	3.00 - 3.49
B-	2.50 - 2.99
C	2.00 - 2.49
C-	1.50 - 1.99
D	1.00 - 1.49
D-	Less than 1.00

Figure 10: Classroom Performance Levels and AMG Ranges

For each level of classroom performance, we compute the average eighth grade test score. We compute score averages separately for minority and majority students, and honors regular course takers. Assuming some correlation between test scores and course grade, we would expect the average exam scores to be highest for the students with the highest course grades. A plot of the data by course and ethnicity (**Error! Reference source not found.**) indicates that this assertion is somewhat true. In all cases, the plot lines slope downward, meaning that as course grade decreases, the exam score tends to decrease as well.

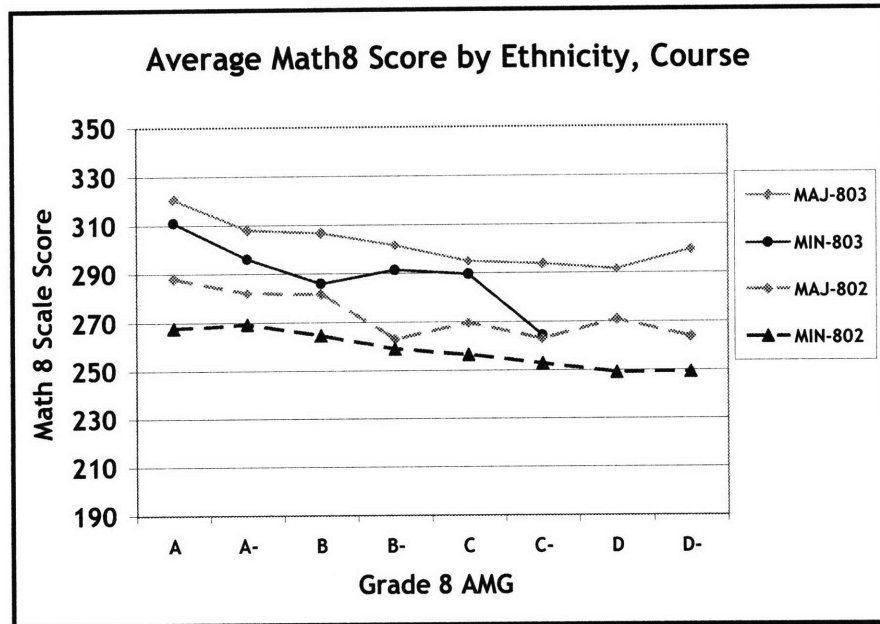


Figure 11: Average *Math8* score by 8th Grade AMG, Course (All Cohorts)

The graph also demonstrates significant differences in the performance of Math 802 and Math 803 students. Within ethnic groups, it is clear that Math 803 students (solid lines) are performing much better than Math 802 students (dashed lines) on the exam. To illustrate using the majority group, students with the lowest course grades in the honors course still scored higher, on average, than their peers with the highest course grades in the regular course.

Within courses, the data show that, at every level of classroom performance, the average ISAT score for minorities is lower than the majority group average. If we assume that test scores and classroom grades are unbiased with respect to race, then there is no obvious reason to expect minorities would underperform their classmates, particularly after receiving similar grades in the course. This finding signals that something unrelated to academic skills may contribute to the testing gap. As mentioned in **Chapter 4.4**, a potential factor would be stereotype threat, if Minority students are concerned about not performing as well as the Majority group, then the testing environment might make Minority students less comfortable in a testing situation, inhibiting their performance.

Although we do not know the cause of testing gaps within grade levels, what seems clear is that, regardless of course grade or ethnicity, honors students perform better on the exam than the students in the regular class. This finding is intuitive, if we assume that the honors kids are simply smarter than their peers. However, the degree of dominance invites the possibility that exposure to the honors classroom environment may contribute to higher test scores.

7.3.2.2. Achievement Gaps within Math Courses

With the knowledge that there were two eighth grade Math courses, we revisit the analysis of **Chapter 4** and examine the eighth Grade achievement gap within the regular (Math 802) and honors (Math 803) Math classes. We summarize the key metrics for both classes in Figure 12; as the results indicate, in terms of grades, there is no achievement gap in the honors course. Average Math grades for minority and majority students are identical, and in absolute and relative terms, there is only a minor difference in outcomes. Recalling that our measures can take on a range of $[-1, 1]$, differences of this magnitude are arguably negligible.

For students in the Math 802 class, the minority student AMG, on average, is slightly higher than a C-plus. For the majority group, the average is four-tenths of a letter grade higher, indicating a slight edge in classroom performance. For both ethnic groups, the mean AMG for Math 802 students is noticeably lower than the mean AMG for Math 803 students. Although it is not possible to determine contributing factors from this analysis, we believe that understanding the reasons why has important implications for the district. On one hand, the mean AMG for honors students may be higher because honors students, as a whole, have better study habits and complete more of their work. On the other hand, teachers may be more inclined to apply a different grading methodology due to different pedagogy or expectations.

Comparison of Performance Metrics Grade 8 Math GPA (2004-2006)		
	Math 802	Math 803
Mean (Minority)	2.4	3.3
Mean (Majority)	2.8	3.3
Rank Sum Index	-0.23	-0.08
Mean Gap	-0.08	-0.02

Figure 12: Grade 8 Classroom Performance by Course (2004-2006 Cohorts)

Observation of testing gaps within the honors and regular eighth grade courses (Figure 13) reveals that the test score gap *among students who take the same math course* is substantially smaller than the overall test score gap. Although there is a difference of thirty scale points in the overall mean performance of majority and minority students, the scale score differences for Math 802 and Math 803 students are sixteen and fifteen points, respectively. The overall testing gap is larger than the gaps in the individual Math courses because of the decidedly different racial composition of the two courses; recall that most of the minority students enrolled in the relatively low scoring Math 802 course and that most of the majority group enrolled in the honors Math 803 course.

Comparison of Performance Metrics Grade 8 Math ISAT (2004-2006)			
	Math 802	Math 803	Math 802+803
Mean (Minority)	258	294	267
Mean (Majority)	274	309	298
Rank Sum Index	-0.41	-0.33	-0.57
Mean Gap	-0.11	-0.10	-0.22

Figure 13: Grade 8 Test Performance by Course (2004-2006 Cohorts)

Although the absolute minority gap is about the same for Math 802 and Math 803 students, the relative measures indicate that there is slightly less overlap in the score distribution of

majority and minority Math 802 students. Judging from test performance, the achievement gaps within Math 802 and Math 803 appear identical. This contrasts with the comparison of AMG data (Figure 12), which indicated a wider gap in Math 802. Notably, despite the absence of a classroom grades gap in Math 803, the testing gap remains.

7.3.3. Grade 8 Placement Rates.

In the previous section, we saw that, regardless of ethnicity, students in the honors eighth grade Math class outperformed their peers in the regular Math class on the eighth grade Mathematics exam. Generally speaking, student enrollment in the honors eighth grade course, or any honors course, is subject to the approval of the student's previous instructor⁵². Presumably, recommendations for honors course are reflective of classroom performance; however, the subjectivity of the process may allow for racial differences in the selection process.

In what follows, we check for consistency between student enrollments in the honors class, and student performance in the classroom. Specifically, we look at eighth grade honors enrollment as a function of sixth grade classroom grades. For a student, the assessment of the sixth grade teacher is the main determinant of whether the student will enter the honors curriculum, and if a student is in the honors seventh grade course, there is a high likelihood that he or she will also take the eighth grade honors course

In our analysis of honors placement, we use empirical data from the 2005 and 2006 cohorts. We have to omit data from the 2004 cohort because in the year that they attended sixth grade, the Oak Park middle schools were for students in seventh and eighth grade only⁵³. We group students according to their Grade 6 AMG, and then calculate the

⁵² According to the district leadership, parents will occasionally influence student placement, but usually teacher recommendations are the ultimate determinant of student placement.

⁵³ For several years, the district has consisted of eight elementary schools and two middle schools; however, prior to Fall 2002, students remained at the elementary school through Grade 6, and attended middle school for two years (Grades 7 and 8) instead of three. As a result, we do not have the same level of detail regarding 6th grade data for the 2004 cohort that we have for the 2005 and 2006 cohorts.

percentage of students who went on to take the eighth grade honors course. We expect that students in the majority group will, on average, have higher course grades in the sixth grade than their minority counterparts. Such a finding would be consistent with the presence of an achievement gap, and also with the disproportionately low representation of minorities in the honors classes. However, among students with similar sixth grade AMG, we would expect majority and minority students to attain honors placement with equal likelihood.

7.3.3.1. Distribution of Grade 6 AMG

The comparison of sixth grade outcomes (Figure 14) confirms a significant difference in sixth grade classroom performance across ethnic groups. In the figure below, we group the students according to their Grade 6 AMG; the lighter bars represent the majority group, and the darker bars represent the minority group. The distribution of majority groups performance is skewed toward the higher grades; over eighty percent of the students had an average grade of “B” or higher. In contrast, minority group performance shows a more even distribution, as the proportion of minorities in the four highest AMG categories is roughly the same as the proportion of minorities in the bottom four AMG categories (52 versus 48 percent, respectively).

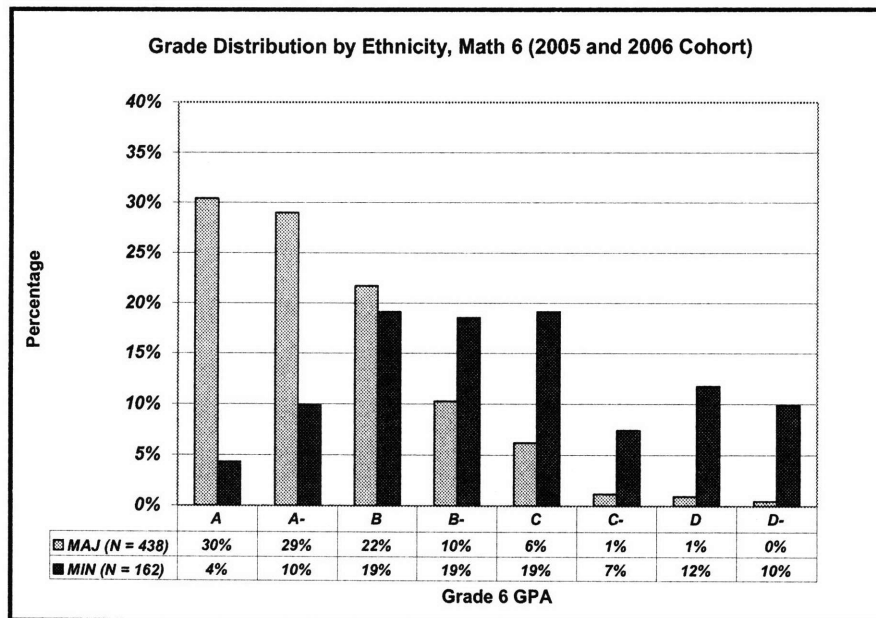


Figure 14: Distribution of for Grade 6 Math Course (2005 and 2006 Cohorts)

The difference in sixth grade classroom performance does much to explain the relatively low participation of minority students in honors courses. Earlier, we observed that over seventy percent of majority group students enrolled in honors courses, as opposed to less than thirty percent of minority students. If the district were to institute a policy that automatically gave honors placement to students with course averages of “B” or higher, then the data indicate that the policy would result in placement rates very similar to the actual result; 81 percent of the majority students would receive honors placement, as opposed to 33 percent of the minority group;

However, in the interest of assessing strategies to narrow the achievement gap, it is still important to compare the honors placement rates of high performing students. Placement in the honors course is a key factor in success on the ISAT exam. Therefore, it would be in the interest of the district to acknowledge and address any barriers that exist regarding minority enrollment in honors courses. Academic skill is an obvious barrier, but by controlling for differences in classroom grades, we may uncover additional roadblocks to honors enrollment.

In Figure 15, we compare the honors (Math 803) placement rates for the Majority and Minority Groups. We are interested in comparing placement rates from a period in time when all students are evaluated using common criteria, before there is an opportunity to split into honors and regular curricula. Thus, we compare placement rates among students on the basis of their average classroom performance in 6th grade. In doing so, our assumption is that the 6th grade classroom performance is a leading factor in the initial decision to place a student in an honors course⁵⁴. For both ethnic groups, we compute the percentage of students in each grade category who went on to take the honors math course, Math 803. For example, the first pair of bars indicates that nearly all of the students with “A” averages in the sixth grade went on to the honors eighth-grade course (99% and 100% for Majority and Minority students, respectively).

For both groups, placement rates decline with student course grade. However, for high performing minorities, the decline in placement has been more drastic. Although “A” students from both ethnic groups enroll in the eighth grade honors course at near parity, minority students in the “A-minus” were less likely to attend (94% vs 81%). For students in the “B” grade range, the difference in placement rates was even larger: whereas 58% of the majority students in this category enrolled in the eighth grade honors course, only 36% of the minority students made it to the honors course.

⁵⁴ Alternatively, a student’s 5th grade ISAT score may also influence the honors placement decision. We create categories of ISAT scores by separating the data into deciles; placement rates in the Majority group decline monotonically with ISAT deciles, with 100% placement in the top decile, and 20% placement in the bottom decile. In the Minority group, ISAT scores are far less reliable indicators of honors placement; only 76% of the Minorities in the top decile received honors placements, and placement rates fluctuated considerably. Ultimately, an analysis based on 5th grade test scores does not alter our conclusions.

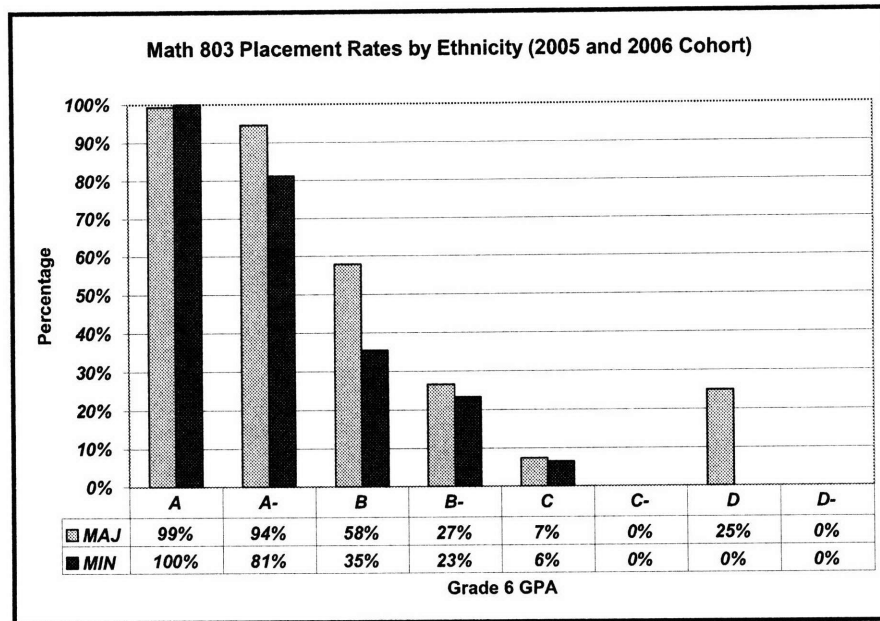


Figure 15: Math 803 Placement Rates by Ethnicity (2005 and 2006 Cohorts)

A graph of seventh grade honors placement (Figure 16) appears to show that the discrepancy in eighth grade placement is a direct result of discrepancies in the seventh grade placement. The difference in Math 703 placement rates among “B” and “A-” students is large as the difference in Math 803 placement. Furthermore, given Grade 6 course grade, placement rates for Math 703 and Math 803 are basically the same for both groups. This indicates that minorities with high course grades are under-represented in the honors Math 703 course enrollment, and that lack of representation carries over into Math 803.

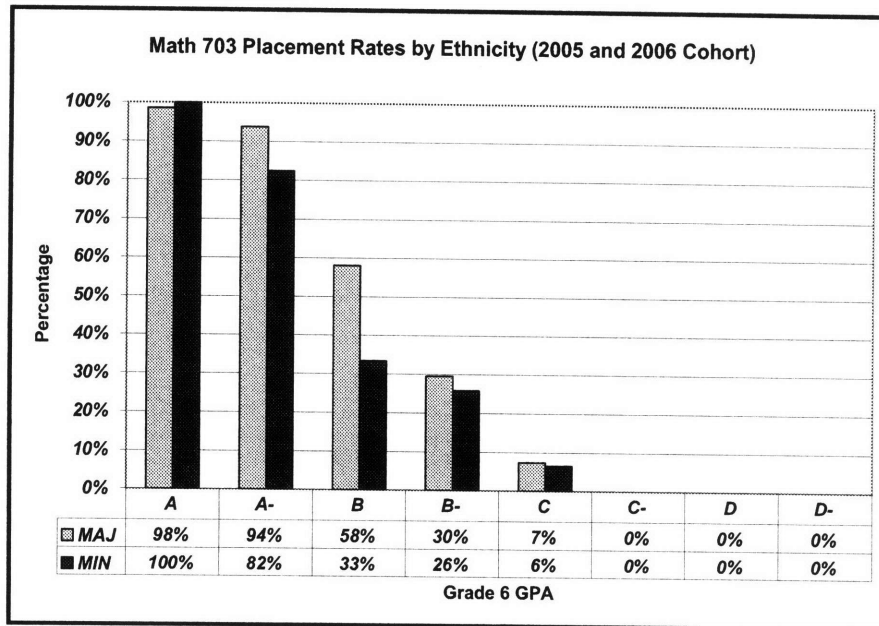


Figure 16: Math 703 Placement Rates by Ethnicity (2005 and 2006 Cohorts)

Given that students in the honors classes score higher on the eighth grade ISAT exam, this discrepancy in student honors placement could potentially contribute to the growth of the achievement gap in Oak Park. On one hand, this trend may simply reflect long-standing differences in skill among honors students and non-honors students. On the other hand, it is also conceivable that exposure to an advanced curriculum and, perhaps, different teacher expectations may provide honors students with a distinct advantage when it comes to test performance.

Under this assumption, if there were no difference in the placement rates of high achievement students, then minority enrollment in honors course would increase. Conceivably, honors placement would improve the test scores of those minority students, thereby narrowing the overall testing gap. In what follows, we develop a model for relating this scenario to the 8th grade testing gap.

7.3.4. Equal Honors Placement and the 8th Grade Math Gap

In this section, we estimate the predictive power of increased minority placement in honors classes on the Grade 8 achievement gap. Our estimations build from two of our previous findings:

- 1) Students in honors courses attain higher scores on the Grade 8 achievement exam than their counterparts in non-honors courses. This finding holds for majority and minority students.
- 2) Minorities who had a “B” or “A-“ average after the first year of middle school were less likely than majority students to enroll in honors courses in subsequent years.

From this evidence, we model a scenario in which honors placement rates for high performing minority students (i.e., students with “B” averages of higher) are the same as those for their White and Asian counterparts. The model indicates that, had this scenario occurred, there would have been a total of nine additional minority students in the Grade 8 honors course; two students with “A-“ averages in the sixth grade math course, and seven with a “B” average.

	Grade 6 GPA Range		
	A	A-	B
Total Minorities	7	16	31
Honors Pct. - MIN Group	100%	81%	35%
Honors Pct. - MAJ Group	99%	94%	58%
Est. Chg. - MIN Group	0	2	7
Difference in MIN Avgs. (Math 803 - Math 802)	NA	33.8	14.1

Figure 17: Estimated Increase in Minority Honors Enrollment

When we control for Grade 6 classroom performance (Figure 18), the data indicate that minority students who attended Math 803 had higher test scores than Math 802 students. For example, among minority students with an “A-“ average in the sixth grade, those who took Math 802 averaged 33.8 points lower than those who took Math 803. Our model for

improvement assumes that, the nine additional minority students would have performed as well as their counterparts in the honors class. Thus, for the students with “B” averages in the sixth grade, we estimate their potential gain as the difference in mean score between Math 803 students who earned “B” averages in sixth grade, and Math 802 students with “B” averages in the sixth grade.

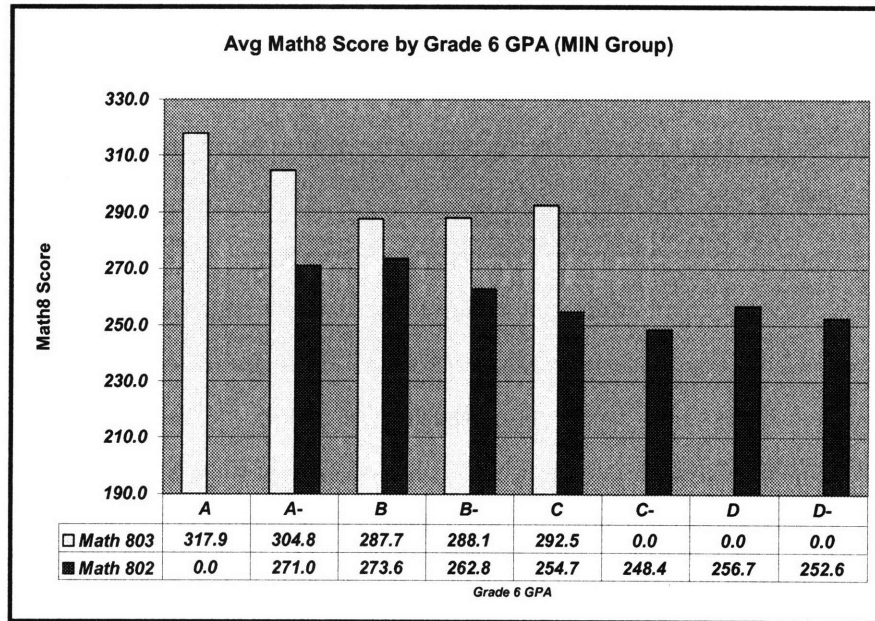


Figure 18: Math8 exam performance, by Grade 8 Math course and Grade 6 course grade

Under these assumptions, the net effect on the Grade 8 score gap is rather small. Although the difference in honors placement rates is significant, our analysis indicates that putting minority transitions into the honors course on par with the majority group would reduce the overall eighth grade gap by approximately 3 percent; see the figure below.

Scenario: MIN placement rates match MAJ placement rates for students with "B" averages or better				
	Grade 6 GPA Range			
	A	A-	B	
Total Minorities	7	16	31	
Honors Pct. - MIN Group	100%	81%	35%	
Honors Pct. - MAJ Group	99%	94%	58%	
Est. Chg. - MIN Group	0	2	7	
Difference in MIN Avgs. (Math 803 - Math 802)	NA	33.8	14.1	
			Sum	
Chng. in cumulative score	0.0	71.7	98.1	169.8
No. of minority students				162
Chng in avg. score				1.0
MIN Avg (Math 802 + Math 803)				269.0
Revised MIN Avg				270.0
MAJ Avg (Math 802 + Math 803)				299.9
Reduction in Score Gap				3%

Figure 19: Gap Reduction Model - Improved Minority Placement in Honors Math

The primary implication is that the differences in group performance are substantial before the students reach middle school. Classroom performance in the sixth grade is a main determinant of whether students attend honors or regular math courses, and the evidence indicates that the lack of minority representation in honors courses is largely explained by the classroom performance gap in Grade 6.

In grades 7 and 8, students are divided into honors and regular math sections. Test performance gaps exist within these math sections, but are roughly half the size of the overall gap, in absolute terms. In terms of classroom performance, majority students in the regular class have a higher average course grade than minority students; in contrast, course grades among honors students indicate that there is no racial gap in classroom performance.

High-performing minority students are less likely to enroll in honors courses than majority students. But although honors placement appears to be a major factor in improving individual outcomes, the relatively low numbers of high performing minority sixth-graders

suggest that this intervention would not predict large changes in the district-wide scoring gap. Next, we explore the role of teacher effectiveness in predicting achievement gaps.

7.4. TEACHER EFFECTIVENESS

The previous section provided an analysis of how course grades and achievement trends vary with respect to ethnicity during the middle school years. Next, we investigate the influence of middle school teachers on the OP97 achievement gap. Our goal is to understand what relationship, if any, exists between a student's progress from Grade 5 to Grade 8, and the teachers that worked with that student.

We anticipate that this analysis will be of special interest to the district. Naturally, the district would like to find better strategies for improving the performance of its students. With this goal in mind, identifying teachers that correlate with high achievement and a low gap could lead to a source of pedagogical best practices for the rest of the district.

Conversely, we may find that particular teachers are notably ineffective in closing the gap. There is some concern, particularly among parents, that some teachers are biased against minority students; if that were the case, we may see evidence in the data. Of course, we would not assume teacher bias in any case, but we expect that the analysis would provide a data-driven perspective on the debate.

7.4.1. Background

There is little doubt that teachers can have a significant impact on student performance. Accordingly, there is a large body of research concerning teacher quality and its relation to student performance. For example, a study of student test scores and teacher history in Tennessee provided evidence that teacher effects are both additive and cumulative in predicting test performance, meaning that a teacher's influence, good or bad, can extend years beyond the initial classroom interaction. Subsequent studies by Wright, et al. (1997)

and Wenglinsky (2002) have argued that teacher quality is the most important school-based factor in predicting educational achievement. With respect to racial achievement gaps, Ferguson (1998) has found evidence that a teacher's perception of a student's abilities is influenced by race, which may result in setting lower expectations for minority students in the classroom. In short, the research suggests that, in a given setting: 1) some teachers would be more effective at promoting Math gains than others; and that 2) a teacher's effectiveness could vary depending on the ethnicity of the student. If there is evidence of either of these patterns among Oak Park teachers, then the district would benefit from knowing that fact.

7.4.2. Methods

In keeping with our general methodology, we use the available data to assess the role of Oak Park's 8th grade Math teachers in predicting student progress. More to the point, we want to know whether some Math teachers in the district are more effective than others, so we are interested in the predictive power of Math teachers on gains in Math proficiency. Our sample for this analysis includes three successive cohorts of student data, and our measure the gain in Math proficiency as the difference in the Grade 5 and Grade 8 ISAT Math score⁵⁵.

Comparing teacher effectiveness on the basis of the performance of their students requires some care, because there are many other factors that play a role in student gains. As we did in our analysis of Reading comprehension and progress in Math, we control for these factors by grouping students based on several common characteristics. For a given grade level, we group students according to their ethnicity (Majority or Minority), the Math course they've enrolled in (regular or honors), and their performance on the 5th grade Math exam (broken into deciles). As a result, we create $2 \times 2 \times 10 = 40$ independent "peer

⁵⁵ We acknowledge that a more appropriate measure would be the one-year difference e.g., the difference in a Grade 7 and Grade 8 exam score. The students in our dataset did not take a Grade 7 ISAT exam; however students in future cohorts will have ISAT exams in Grades 3 through 8, inclusive.

groupings” from which to compare relative gains. In what follows, we discuss how we use these groupings to develop a measure of teacher effectiveness.

7.4.2.1. Rating Teacher Effectiveness: An Example

For this example, suppose we are interested in rating teacher effectiveness with respect to the minority students who took the regular eighth grade Math course (Math 802). We derive the measure for a teacher’s effectiveness from the gains made by their students (Math8 – Math5) on the state exam. Suppose there are 100 such students in our sample. As a first step, we separate the data for these students into ten “peer groupings” based on their performance on the Grade 5 exam. Each grouping, by construction, then contains a group of students with reasonably close Grade 5 exam scores⁵⁶.

Now, within each grouping, there will be an assortment of students, taught by 1 or more different instructors. In Figure 20, we have a peer group of ten “Math 802” students (labeled s01- s10), taught by four different instructors (labeled as t07, t14, t33, and t58). The headings in Figure 20 uniquely define the ethnicity (minority group), math course (Math 802), and previous achievement category (Decile 10) of the students in the group. In this example, these are the highest scoring minority students on the Grade 5 exam who were placed in Math 802.

⁵⁶ The number of “peer groupings” formed (10, in our case) can be adjusted if necessary. The purpose of splitting the data is to avoid comparing gains made by students with high scores on the Grade 5 exam from gains made by students with low scores on the Grade 5 exam.

Ethnic Group: Minority				
Math Course: Math 802				
ISAT5 Achievement : Decile 10				
StudentID	Teacher	Point Gain	Rank	Percentile
s01	t33	64	10	0.95
s02	t33	60	9	0.85
s03	t14	57	8	0.75
s04	t33	51	7	0.65
s05	t33	46	5.5	0.50
s06	t33	46	5.5	0.50
s07	t33	43	4	0.35
s08	t07	40	2.5	0.20
s09	t58	40	2.5	0.20
s10	t58	37	1	0.05

Figure 20: Sample Peer Grouping: Grade 5 – Grade 8 ISAT Test Gains

Within this group, we measure the gains made by each student, and rank the gains from highest to lowest. For example, the student with the largest gain has the largest rank (10), the next largest gain has a rank of 9, and so forth⁵⁷. Rather than express the rank in terms of the number of students (e.g., rank k of n students), we convert the rank to a percentile p using the following formula:

$$p = \frac{k - 0.5}{n}, k \leq n$$

These percentiles are calculated independently for each of the deciles; ultimately, every minority student in Math 802 will have an associated percentile, representing his gain in test score, relative to the peer group. For a peer group of size n , the percentile takes on discrete values within the interval $[0.0, 1.0]$.

Now, every minority student who took Math 802 has a percentile score that represents the progress made relative to his or her peers. When we these group these students by teacher, we use the average student percentile as a measure of teacher effectiveness. Figure 21 demonstrates the grouping that would occur for the students in Decile 10. Although a teacher will not necessarily have a student in every decile, when the students are grouped

⁵⁷ In the event of ties, we use fractional ranks; For example, if two students had tied for the largest gain, they would both receive a rank of $(9+10)/2 = 9.5$.

across deciles, each Math 802 teacher will receive an effectiveness rating based on the relative gains made by his or her students.

Ethnic Group: Minority		
Math Course: Math 802		
ISAT5 Achievement : Decile 10		
Teacher	Student(s)	Avg. P'tile
t07	s08	0.20
t14	s03	0.75
t33	s01,s02,s04,s05,s06,s07	0.63
t58	s09,s10	0.13

Figure 21: Teacher Effectiveness Ratings for the Sample Peer Group

To define the rating metric explicitly, let S represent the set of students s that belong to a given ethnic group. Recall that each student's percentile is a measure of the progress made on the Grade 8 exam, relative to a peer group with similar Grade 5 test scores; let $p(s)$ represent the percentile of student s . Now, for every course instructors t ,⁵⁸ we define $S(t)$ as the set of students taught by instructor t , and we define the teacher effectiveness rating, $Q(t)$, as the average relative performance of a teacher's students:

$$Q(t) = \frac{1}{|S(t)|} \sum_{s \in S(t)} p(s)$$

As the average of student percentiles, $Q(t)$ can only take values on the range $[0,1]$, and the expected value of $Q(t)$ is 0.5, which would correspond to the district average for teacher effectiveness. This measure should have some intuitive appeal for, over time, we would expect a teacher with a rating of 0.5 to have an equal number of students with relatively large gains (i.e., percentiles above 0.5) and students with relatively low gains (percentiles below 0.5). As a result, predicting a future student's progress after having a teacher with a 0.5 rating would be analogous to a coin flip between above-average and below-average student progress.

⁵⁸ In our analysis, we elected to only compare the teachers who had taught at least 5 students in S .

However, we must also address the presence of sampling error. Even if there were such a thing as a truly “average” teacher, it is unlikely that we would observe an effectiveness rating of exactly 0.5 in a limited group of students. We address this issue by calculating p-values for each value of $Q(t)$. Under the assumption that the long-term, “true” effectiveness rating would be 0.5, the p-values tell us the likelihood of our observed value $Q(t)$.

The p-values are determined by the value of $Q(t)$, and by the number of observations used to calculate $Q(t)$; that is, the number of students who had instructor t . Because the number of observations can become very small (in some cases, less than 10, we use Monte-Carlo simulation to calculate (approximate) p-values for $Q(t)$ ⁵⁹ ⁶⁰.

Before discussing the findings of this approach, we briefly summarize the steps of our procedure.

1. Separate the minority students into deciles based on Grade 5 math score.
2. Calculate the difference in score (gain) from Grade 5 to Grade 8 and, within each decile, rank the students’ gains, with 1 representing the smallest gain.
3. Convert the ranks into student percentiles. For a student of rank k in a decile containing $n \geq k$ students, the percentile p is defined as follows:

$$p = \frac{k - 0.5}{n}$$

4. For each teacher, find the percentile values for all students who had that teacher and calculate the average, q^* . This is $Q(t)$, the teacher’s effectiveness rating.
5. For each q^* , determine an approximate p-value via Monte-Carlo simulation (10000 instances) regarding the following hypothesis:

H0: Scoring gains do not vary significantly across teachers in Oak Park (i.e., teacher effectiveness = 0.50 for all teachers)

⁵⁹ In our simulation, we model $Q(t)$ as the average of m IID uniform (0,1) random variables, where m is the number of observations.

⁶⁰ The model is approximate because the percentiles which comprise $Q(t)$ are not necessarily independent.

6. The p-value is the likelihood that the hypothesis is true, given the value of q^* . Compute the p-value directly by calculating the proportion of simulated values of q that lie farther than $|q^* - 0.5|$ from the mean of the sample distribution (~ 0.5)
7. Repeat the procedure for students in the majority group

7.4.3. Results

We begin with a review of the teachers who taught Math 802, the regular Grade 8 Math course. Under our methodology, we identify highly effective teachers in the district by looking for values of $Q(t)$ larger than 0.5 and low p-values. Large values of $Q(t)$ denote the teachers whose students, on average, outperformed their peers. Low p-values indicate that the observed effectiveness is not likely due to chance. The presence of both factors does not prove that the teacher is responsible for the relatively strong performance of his or her students. However, it does provide evidence that the teacher's students are making atypical amounts of progress, a trend that warrants further investigation.

7.4.3.1. Teacher Effectiveness – Math 802 Minority Group

Our first effectiveness ratings address the performance of the Math 802 minority group (Figure 22). The list of instructors includes any teacher who taught at least five of the minority students in our dataset; as the table indicates, there were ten teachers who met these criteria. (Teacher names are disguised to preserve confidentiality).

The top row indicates that Teacher #7 ($t07$ in the table) taught 13 of the minority students in our sample, and on average, these students made larger gains than their peers just over half the time, giving Teacher #7 an effectiveness rating of 0.518 on a $[0, 1]$ scale. This rating is a difference of 0.018 from the mean teacher rating (0.5), so there isn't much reason to believe that this teacher is significantly more (or less) effective than the other teachers. This conclusion is supported by the p-value, which indicates that, under the assumption of a neutral relative effectiveness, we might see differences of 0.018 or larger about 80% of the time.

Course 802		MIN Group	
Teacher	Students	Avg. Ptile	Pvalue
t07	13	0.518	0.82
t10	6	0.201	0.01
t14	12	0.618	0.16
t18	55	0.549	0.21
t28	31	0.471	0.58
t33	20	0.380	0.06
t42	8	0.825	0.00
t50	18	0.478	0.76
t58	11	0.426	0.41
t70	12	0.447	0.53

Figure 22: Math 802 Teacher Ratings – Minority Students (2004-2006 Cohorts)

In contrast, Teacher #42 has an effectiveness rating of 0.825, a sign that this teacher’s students are consistently making more progress than their peers. Furthermore, despite the relatively small number of minority students in the sample (8 over three years), the p-value indicates that this difference is not a matter of random error. In a similar vein, the students taught by Teachers #10 and #33 have smaller gains, on average, and the low p-values indicate that these findings are also not due to random error.

7.4.3.2. Teacher Effectiveness – Math 802 Majority Group

Next, we evaluate the Math 802 teachers based on the performance of their majority group students (Figure 23). Again, the students taught by Teacher #42 appear to make much more progress than their similarly situated peers. As we saw with the minority group data, the majority group students taught by Teachers #10 and #33 also made smaller gains than their peers on average. However, the associated p-values are notably larger, an indication that the low ratings are possibly due to sampling error. Among the majority group, the students who had Teacher #70 made the smallest gains. Figure 22 indicates that this teacher’s minority students also performed below average, but the difference wasn’t as substantial.

Course 802		MAJ Group	
Teacher	Students	Avg. Ptile	Pvalue
t07	17	0.551	0.47
t10	7	0.473	0.80
t14	18	0.599	0.15
t18	51	0.483	0.66
t28	46	0.471	0.50
t33	21	0.459	0.52
t42	11	0.882	0.00
t50	7	0.512	0.92
t58	8	0.367	0.20
t70	12	0.284	0.01

Figure 23: Math 802 Teacher Ratings – Majority Students (2004-2006)

7.4.3.3. Teacher Ratings across Ethnic Groups

For the most part, the test gains did not indicate large variances in the effectiveness of the Math 802 faculty. Among teachers, there are clearly some differences in the relative gains made by their minority and majority students. But when we compare the average gains made by both ethnic groups (Figure 24), we see that most teachers rated at or near the “average” effectiveness rating of 0.5.

The data from Math 802 also suggest that teacher effectiveness is largely consistent across ethnic groups. For most of the teachers, the minority and majority average gains are fairly close to each other. For 8 of the 10 teachers, the ratings are within 0.1 of each other. And in 8 of 10 the cases, the ratings are in the same direction from 0.5 for both minority and majority students. The consistencies in ratings suggest that identification of an “above-average” (or “below-average”) teacher does not hinge on student ethnicity.

If we were to assume that teachers do influence student gains, these findings indicate that *the relative effects of teachers remain largely intact, regardless of the student’s ethnicity*. This finding also supports the notion that effective teachers will tend to be effective for students of all races, although external factors may mitigate the degree of a teacher’s predictive influence.

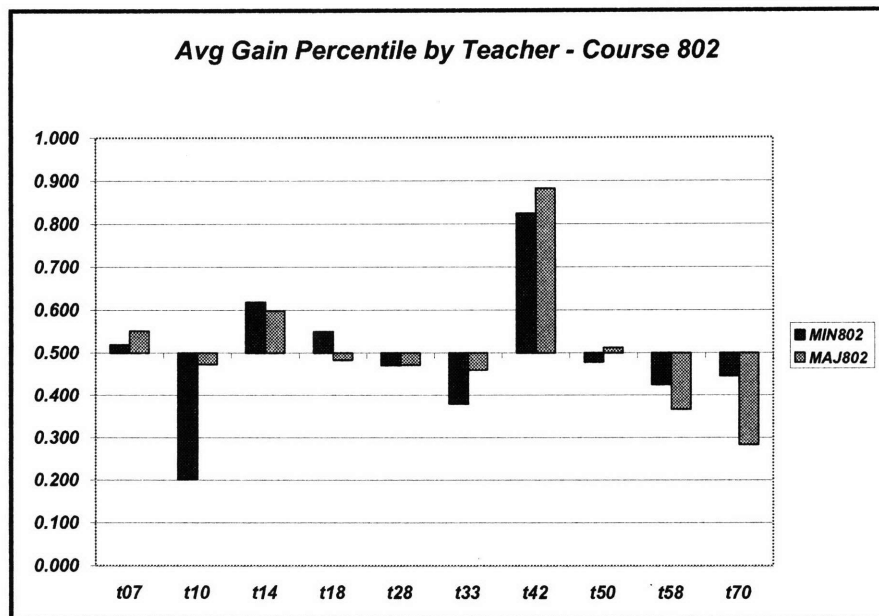


Figure 24: Comparison of Avg. Student Gain by Teacher – Math 802

When there are substantial differences in the minority and majority group gains, there is the obvious concern about teacher practice. In the case of Teacher #10, for example, we see that the minority student rating is much lower than the rating for the majority group students. Teacher #70's ratings reveal the opposite trend, a teacher with a noticeably lower rating for majority group students. These discrepancies might reflect chance fluctuations within limited groups of students, but they might warrant further inquiry.

7.4.3.4. Teacher Effectiveness - Math 803

At this point, we move on to rating teacher effectiveness in the Math 803 course. Many of Oak Park's math instructors teach both the regular and the honors course, but by developing separate ratings, we allow for the possibility that a teacher's effectiveness might vary with the type of course, due to a different approach, expectations, etc.

The tables of Figure 25 list the ratings (minority and majority, respectively) for each Math 803 teacher. The group includes all of the Math 802 instructors, with the addition of an

eleventh instructor, Teacher #51. When we compare these student counts to the Math 802 tables, we see that, only 7 of the 11 Math 803 teachers saw 5 or more minority students in their class. Although we generate ratings for both ethnic groups, we will only be able to make direct comparisons with 7 of the teachers

MIN Group Course 803				MAJ Group Course 803			
Teacher	Students	Avg. Ptile	Pvalue	Teacher	Students	Avg. Ptile	Pvalue
t07				t07	21	0.525	0.68
t10				t10	21	0.637	0.03
t14	5	0.730	0.07	t14	44	0.480	0.65
t18	7	0.427	0.51	t18	111	0.542	0.13
t28	5	0.388	0.40	t28	38	0.509	0.85
t33	24	0.441	0.32	t33	65	0.372	0.00
t42	8	0.659	0.11	t42	75	0.697	0.00
t50				t50	35	0.444	0.25
t51				t51	16	0.497	0.96
t58	9	0.464	0.71	t58	41	0.394	0.02
t70	7	0.557	0.62	t70	37	0.294	0.00

Figure 25: Math 803 Teacher Ratings – Minority and Majority Students (2004-2006)

A plot of the teacher ratings (Figure 26) shows that the students taught by Teacher #42 made larger gains than their peers, regardless of ethnicity. The low number of minority students makes the minority rating less reliable on its own, but the Math 803 ratings are comparable to the ratings derived from the students in Math 802. Again, the minority and majority group ratings tend to be of comparable magnitude, indicating that, if there is a teacher effect, then the direction of the prediction tends to be the same for both ethnic groups. As for exceptions, Teachers #14 and #70 show the largest discrepancies in teacher ratings; in both cases, the majority group average is notably lower.

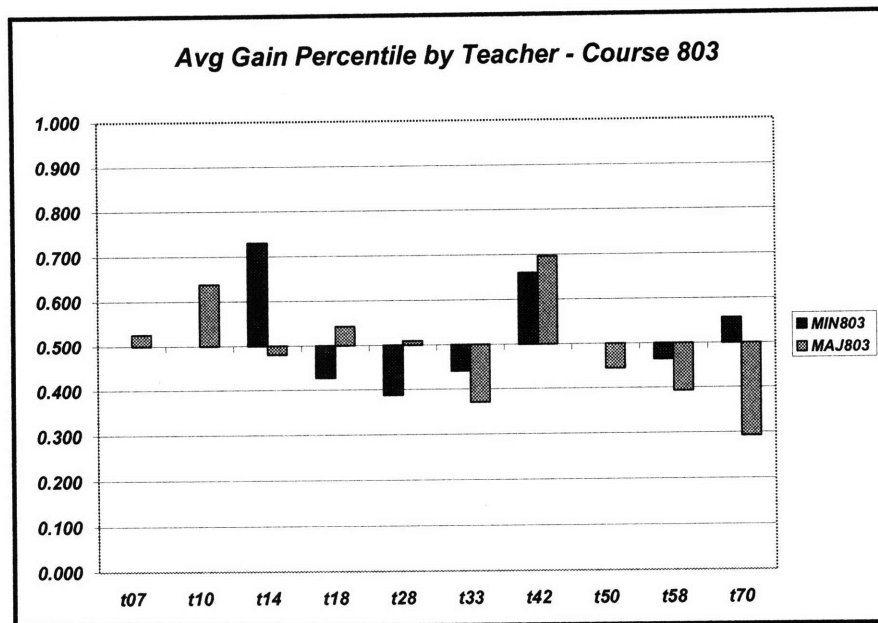


Figure 26: Comparison of Avg. Student Gains by Teacher - Math 803

As mentioned, we create separate ratings to allow for the possibility that a teacher’s effectiveness may vary between the honors and regular course. Interestingly, the teacher identified as Teacher #42 was the only teacher to receive high ratings for both ethnic groups, and for both courses. Teachers with uniformly high ratings are associated with large student gains across a diverse range of students; as such, these teachers might be especially helpful in helping boost minority achievement in the district. However, the predictive relationship to the achievement gap is unclear, for a rise in student achievement for both ethnic groups might merely shift the achievement gap upward.

7.4.4. Other Comments

We close with some comments regarding the use of this metric. The findings here represent an approach to using student data as a measure of relative teacher effectiveness. In doing so, the approach assumes that teachers can influence student test achievement, a notion that has some support in academic literature.

We believe that this metric is a useful complement to existing measures of teacher effectiveness. Within the Oak Park district, there has been some anecdotal evidence to support this claim. When we initially shared these findings with the district, we did so without revealing teacher identities. Upon presentation, the district leaders wanted to know the identity of Teacher #42, the teacher that appeared to be highly effective with all groups of children. After resending the analysis with the names attached, we learned that, independently of our analysis, Teacher #42 had received recognition for their teaching. The case of Teacher #42, while not conclusive, provides some evidence that our rating system can point a district to examples of above average effectiveness among their teachers.

7.5. SUMMARY

In this chapter, we explored the ways in which achievement data might change in response to district initiatives to close the minority Mathematics achievement gap. In doing so, our intent was to offer a data-driven perspective on the initiatives whose data predict large reductions in the gap.

In general, our findings suggest that the district might do well to focus its efforts on early interventions, although there is some opportunity to improve student outcomes in the middle schools; the largest opportunities for improvement appear to be connected to interventions at the elementary school level, or earlier. Our data indicate that gaps in achievement start large and grow larger over time. Students leave the district after eighth grade, and the district's ability to narrow gaps before they leave relates directly to the district's ability to intervene early.

We had observed in the previous chapter that the eighth grade achievement gap is traceable to achievement gaps that show up as early as the third grade. However, we also note that high-performing minorities are less likely to maintain high levels of performance over than those in the majority group. If minorities in the district were able to match the progress made by their peers in the majority, we estimate that the nominal eighth grade achievement

gap would decrease by nearly 25%. We also find evidence of a correlation between reading proficiency, and gains made on the Mathematics exam; under the assumption of causality, if the district were to eliminate the Reading gap by 5th grade, we again estimate a potential 25% reduction in the eighth grade Mathematics gap.

Our examination of the district's middle schools provided important insights into the Oak Park educational experience, highlighting the influence of the course curricula and teachers on student performance. The district offers a regular and an honors level Math course for Grades 7 and 8; within our sample, less than 30% of the minority students had placement in an Honors course, as opposed to about 70% of the majority group. After sixth grade, instructors determine which students receive Honors placement, presumably on the basis of classroom performance; although there is evidence that minority sixth graders with high math grades were less likely to receive Honors placement, there were so few high performing minorities that discrepancies in Honors placement predict less than 10 percent of the 8th grade gap.

Among the teachers of Oak Park, we used student gains on the Math exam to construct a measure of relative teacher effectiveness. We create separate effectiveness ratings, based on student ethnicity (Minority or Majority), and the Math course taught (Regular and Honors). In most cases, a teacher's effectiveness was consistent across races; a teacher's relative effectiveness did not vary significantly between majority and minority students. Also, most teachers demonstrated comparable levels of overall effectiveness. Despite the general trend, there were exceptional instances in which a teacher stood out, due to above average student gains across ethnic groups or due to large variations in effectiveness between ethnic groups.

The results of the preceding chapters provide insights into the nature of the Oak Park achievement gap, using the data made available by the district. That said, our findings are subject to a number of important caveats. In the remaining chapters, we briefly review those caveats before concluding with a general summary of all that we have observed.

8. Caveats

In this thesis, we have explored a variety of issues regarding achievement and race in Oak Park. The results and implications of this work are subject to a number of guiding assumptions. Before we conclude, we would to remind the reader of a few very important things to bear in mind regarding this study. All of these caveats should be familiar, but we review them here to ensure that the results of our work remain in the proper context.

- *The findings reflect the trends observed in a single and somewhat atypical elementary school district.*

It is important to remember that the findings presented herein refer to a specific school district; namely, the Oak Park elementary school district. The Oak Park community is atypical in some respect, with middle-class income, substantial racial diversity, and a relatively extensive dataset. The achievement gap metrics and the various statistical tests and methods used to study the Oak Park district have the potential for use in the study of other school districts. However, the findings regarding the magnitude and evolution of achievement gaps, as well as the correlations with other school factors might vary considerably in another district.

- *Many of the findings rely on state test data.*

Several of our analyses employ standardized test data as the measure of student achievement, and by extension, as a proxy for student proficiency. However, there is reason to believe that the state exam may overstate gaps in proficiency. In general, although standardized testing is the *de facto* means of measuring student achievement at

the state and national level, there are legitimate concerns regarding the accuracy of standardized tests as a measure of student proficiency. In this thesis, we have mentioned the possibility of increased test anxiety among black students due to stereotype threat (**Chapter 2**).

Where possible, our analyses were repeated using classroom grades as the achievement measure. When comparing testing gaps and grade gaps, we found that testing gaps were frequently the larger or the two. Further, when we control for alternative measure of achievement, we have seen ISAT gaps persist. This occurred when we compared to 3rd grade ISAT data among students with comparable 3rd grade Stanford 9 exam data (**Chapter 2**) and again when we compared 8th grade ISAT data among students with comparable 8th grade math grades (**Chapter 7**).

Regarding our use of testing data, we assume that the testing instruments are fair and reliable indicators of student proficiency. The Illinois State Board of Education (ISBE) has found the cut points for performance levels and the method for calculating scale scores to be acceptable, and we do not challenge the point. Also, as suggested by ISBE, we assume that the pre-2006 version of the ISAT is indeed of comparable rigor to the new ISAT exam, and that either exam is comparable to the Stanford Tests (series 9 and 10).

- *The findings reflect the perspective of quantitative data, tracked by the district.*

The scope of this study is limited to a set of readily available and easily quantifiable school and environmental factors and academic outputs. The use of readily available data is important, because it permits us to study achievement in terms of the district's own data. Accordingly, we focus on the synthesis of existing data, and there were no attempts made to collect new data from the students or from the member schools of District 97.

However, because of our rather specific focus, this analysis does not consider a number of other factors and outputs, either because they not readily tracked by the district, or because they are less quantifiable, or both. Although these findings portray the achievement gap from the perspective of district data, we do not claim that this perspective is definitive. For example, in our analysis of honors course placements (**Chapter 7**) we assume that, if high-performing minority students attend in honors courses, they will perform as well as the minority students already there. However, this is not a given; as noted by Ferguson (2001), differences in a student's motivation may preclude entry into honors classes, much less higher scores.

- *The correlation of student factors and achievement outcomes does not imply causation.*

In the presence of correlation, our findings in **Chapter 7** are contingent on certain assumed directions of causality. The connection could be causal (e.g., “Reading comprehension improves Math performance”), or complementary (e.g., “Factors that improve Reading comprehension also improve Math performance.”); however, if neither assumption is true and no “true” relationship exists, then we have no basis for assuming our assumptions would still hold.

To illustrate, we note that children with better reading scores tend make more progress in Math between 3rd and 8th grade. Our estimates of narrowing the math gap by improving Reading skills are contingent on the notion that better Reading skills will lead to bigger Math gains. However, better Reading skills are not necessarily the determining factor; for instance, the key factor might be an attitude towards studying that improves Reading and Math skills.

Even if the causal assumptions are correct, it isn't clear how the school district should act on them. In the Reading and Math analysis, we estimate that closing the Reading gap would cut the Math gap by about 25%. For all intents of purpose, we assume that the

eliminating the reading gap might be just as difficult as ending the math gap directly. The intent of our analysis is to determine the most promising strategies; however, the district must always consider the size of the opportunity in the context of what is feasible.

9. Conclusions and Future Work

The study reflects our attempts to accomplish a number of objectives. The first was to survey current approaches to measuring achievement gaps, and to propose complementary methods, if need be. The second objective was to identify patterns of math achievement within different ethnic groups in Oak Park. The third objective was to use district data to detect the prevalence of several key factors believed to influence the minority gap in Oak Park. Under the assumption that a factor was influential, we then sought to develop models to estimate the reduction in the minority gap that might occur in the absence of that influence. Here, we summarize our findings and conclusions with regard to each of these objectives.

9.1. MEASUREMENT

Our assessment of current metrics for measuring achievement (**Chapter 3**) indicates that there are notable limitations among the most popular approaches. Benchmark metrics, which are widely used by schools and states alike, can obscure substantial variations in performance. Normalized (“z-score”) measures, which are commonly seen in academic literature, require implicit assumptions about data distribution and sample size that are not necessarily valid.

As we consider alternative approaches to measuring achievement gaps, we also note that one’s perception of achievement trends may change depending on how achievement gaps are measured. We consider the various approaches into a mix of relative and absolute

measures. To help define the dichotomy, relative measures focus on “who’s ahead”, whereas absolute metrics consider “who’s ahead, and by how much.”

In light of these findings, we believe that the prospect of identifying a single, “correct” way to measure achievement gaps is unlikely. If the objective is to provide a summary of group outcomes, than choosing the best metric is a matter of taste. However, if the goal is to use achievement data as an analytical tool (e.g., to isolate trends in performance), educators might do well to use a variety of relative and absolute measures to study student performance.

To this end, we employed a collection of achievement metrics designed to measure achievement gaps in absolute and relative terms. For consistency, each of the metrics are calibrated on a [-1, 1] scale. We calibrated the scale so that negative values indicate that the minority group, which is mostly Black, has not performed as the majority group, which is mostly White. A value of zero implies that there is no evidence that either group is ahead.

9.2. ACHIEVEMENT GAP DYNAMICS

Our analysis of exam results within Oak Park (**Chapters 4 and 5**) confirms the existence of a persistent and substantial gap in the performance of minority and majority students. Such a gap arose in every analysis we performed, at every grade level in every cohort with every metric we used. This achievement gap appears larger in relative terms than in absolute terms. This is partially because, when compared to other elementary school districts in Illinois, Oak Park students tend to outperform their peers on the state mathematics exam, regardless of race. Minority students in Oak Park are only underperforming relative to majority students in the district, not to all students in Illinois.

Using the 8th grade class of 2005 as a baseline, the data reveal a performance gap on the 8th Grade Mathematics exam and, using course grade data, in the classroom as well. Course grades are a valuable complement to exam scores because they capture the teacher's assessment of student performance. The achievement gap is smaller in magnitude when measured using course grades, which leave room for the possibility that standardized tests, as a measure of performance, may overstate the actual gaps in student knowledge. However, the classroom gap is still large, as the average Math grade (AMG) for a minority student (2.3 on a 4.0 scale) was nearly a full letter grade lower than the majority group average (3.0).

Across three cohorts of Oak Park students, we observe the achievement gap observed in 8th grade begins far earlier than that.. Test data among students who attend the district for several years indicate that the testing gaps attain most of their magnitude as early as 2nd grade. This is true in absolute and relative terms. In terms of the learning standards, most students show the same level of proficiency (above, at, or below standards) in the eighth grade as they do in the third grade, regardless of race. In aggregate, this trend appears to support the notion that achievement gaps remain static throughout elementary school; however, among students who exhibit comparable levels of performance on the 3rd grade Math exam, there is statistically significant evidence that minority students consistently made less progress over time than their peers in the majority group.

9.3. POTENTIAL CORRELATES OF ACHIEVEMENT GAPS

Chapters 6 and 7 of the thesis chronicle a series of analyses aimed at investigating factors believed to correlate with the Oak Park achievement gap. The district selected each of the topics covered in this study, thus, our findings directly address elements that were of importance to the district.

9.3.1. Low-Income and Student Mobility

Although minority students are more likely to live in low-income households, our analysis indicates that differences in income explain little of the sustained difference in majority/minority outcomes in this district. School data indicate that minority students are far more likely to be eligible for free/reduced lunch programs. However, the achievement among students not receiving the subsidy is just as large as the overall gap. Despite the correlation between low income and low achievement, academic performance among minorities does not generally improve with income, and thus, the achievement gap exists across all income groups.

Regardless of the strategies employed at Oak Park, the high mobility of minorities will likely hamper the progress made in closing the gap. Of the students who graduated from 2004 through 2006, only 50 percent of the minority students were veteran Oak Park students. Under the assumption that Oak Park can find strategies for accelerating the performance of its minority students, a significant percentage of students will transfer in or out of the district before incurring the full benefit of the intervention. It is important to remember that the results of our study reflect observations drawn from the so-called “veteran” students of Oak Park, the ones who graduate after having spent at least five years in the district.

9.3.2. Achievement and Gender

Our study of gender effects on student achievement indicates that females demonstrated more progress in math over Grades 3 through 8 than males, regardless of race. Minority males and females exhibit comparable levels of performance in third grade, but as females make more progress, minority males exhibit the lowest levels of achievement on the eighth grade exam. Among the majority group, males actually outperform females in third grade, but the females catch up to the males over time and the two exhibit nearly identical levels of performance on the eighth grade test.

We examined the transitions in student Math performance from Grade 3 to Grade 8 as a measure of district effectiveness⁶¹. By this measure, the district achieved its highest levels of effectiveness with females in the majority group, followed by majority males, minority females and minority males, in that order. Had the district been as effective with minority students as it was with majority students during Grades 3 through 8, we estimate a reduction in the 8th grade race gap would of roughly 25 percent. That outcome implies that a large portion of the gap (75%) would still remain, due to the sizable ethnic gap that exists by third grade. If the district had been as effective with minority males as it had been with minority females, the gap between minority and majority students would have decreased by approximately 7 percent. The decrease is rather modest, indicating that, when compared to the performance of the majority group, the performance of minority males in the district is not that dissimilar from the performance of minority females after all.

9.3.3. Achievement and Reading Comprehension

Our study shows strong evidence of a positive relationship between a student's 5th grade reading proficiency and the gain on the math exam made between Grades 3 and 8. Controlling for gender, ethnicity and third grade math performance, we found that in 35 of 40 independent samples, there was a positive correlation between reading performance and math gain. Across samples, a 10-point improvement in the Grade 5 reading score would equate to, on average, a 3-point increase in the progress in math made from Grade 3 to Grade 8.

On the basis of this relationship, the data suggest that, had there been no 5th grade reading gap, there might have been a 25 percent reduction in the 8th grade Math gap. This finding does not presume that closing the reading gap is any less challenging than closing the math gap in Oak Park. However, this finding does suggest that efforts to improve reading skills,

⁶¹ This definition of effectiveness explicitly takes into account the possibility that students who are different levels of performance in third grade might progress at different rates.

particularly during the elementary school years, might also lead to improved math performance.

9.3.4. Achievement and Honors Placement

At the middle school level, the district offers Standard and Honors Math courses for Grades 7 and 8. In general, honors students outperform students in the standard Math course on the 8th grade exam. This is true regardless of racial background; despite the district wide achievement gap, which favors the majority group, the minority students in the honors course outperform the non-minorities in the regular course. However, these students are relatively few in number. From our sample, over 70 percent of majority students were enrolled in the eighth grade Honors course. In contrast, fewer than 30 percent of the minorities took the eighth grade Honors course.

Racial achievement gaps are smaller in magnitude within the Standard and Honors Math courses. When compared to the overall testing gap, absolute gaps on the 8th grade state-wide exam *within* each course are half as large. In relative terms, the gap is two-thirds smaller in the standard Math course, and half the size in the honors course. Comparing course grades, majority students in the Standard course had a mean classroom grade (AMG) of 2.8, and minority students had a mean AMG of 2.4. In the honors course, the mean AMG for majority and minority students is 3.3, an indication that there is no achievement gap in grades among Honors students.

Honors placements reflect, in part, an assessment from the 6th grade Math teacher and honors enrollment is highly correlated with high grades in the Grade 6 Math course. There is evidence to suggest that minorities who do well in the Grade 6 math course are less likely to enroll in honors classes than majority group students. However, our calculations suggest that increasing the enrollment of high-achieving minorities in honors courses would decrease the overall testing gap by less than 5 percent. This is because the number of minorities with grades of B+ or higher in the 6th grade is quite small. This low number of

high-achieving minorities is another indication of substantial differences in math performance before the students reach middle school.

9.3.5. Achievement and Teachers

Under the assumption that test score gains can represent teacher effectiveness, our findings suggest that certain teachers do seem to be associated with relatively large gains in student performance. Conversely, other teachers are associated with relatively small test score gains. Teacher effectiveness can vary, depending on the type of math course taught. However, there is little evidence of racial bias, as teachers tend to demonstrate comparable levels of effectiveness with minority and majority students. Put simply, good teachers do well with all students, regardless of ethnicity.

9.4. COMMENTS

Although achievement gaps in education exist on a national scale, our study supports the notion that the study of achievement gaps at the local level are an essential area for future research. Our study also corroborates the existence of substantial achievement gaps that extend beyond comparisons between inner cities and affluent suburbs, all the way to ethnically diverse, middle-class districts like Oak Park.

Our work with Oak Park has shown that achievement data can be used to study local achievement gaps and assess the potential benefits of pursuing a particular strategy. As the district builds a richer set of student data, the indications may very well change. Since the gathering of our data, student testing has expanded to annual exams between Grades 3 and 8, inclusive. The availability of more data will presumably improve the quality of the analysis, enabling the district to refine its analysis of changes in the district.

Moving forward, we offer some final thoughts on achievement gaps. First, we would recommend that the district continue to build the dataset for its student body. Ideally, the district would maintain a record of teachers and course grades for every year that a student is in the district. This information has been maintained for several years at the middle school level, but our study indicates that this data may be more useful if also available at an earlier stage in school.

Unfortunately, the data thus far does not reveal any easy answers to closing achievement gaps. Even if all the measures we considered to close the ethnic Math gap were undertaken and were maximally successful, a substantial achievement gap would remain in Oak Park. Perhaps that outcome means that eliminating the gap is too much to hope for, except over a time frame far longer than the foreseeable future. In lieu of discovering the correct strategy for erasing achievement gaps, districts like Oak Park would do well to assess the potential of their strategies, and learn what they can from their local data.

10. References

- [ARI07] L Aratani. "Finding Ways to Better School African American Boys" *Washington Post* (May 17, 2007).
- [BAL04] VA Bali and RM Alvarez. "The Race Gap in Student Achievement Scores: Longitudinal Evidence from a Racially Diverse School District," *The Policy Studies Journal* 32:3 (2004), 393-415.
- [BOU05] KP Boudett, EA City, and RJ Murnane (eds.); *Data Wise*. Harvard Education Press (2005).
- [CAI70] GG Cain and HW Watts. "Problems in Making Policy Inferences from the Coleman Report." *American Sociological Review*, Vol. 35: 2. (Apr. 1970): 228-242.
- [CLO06a] CT Clotfelter, HF Ladd, and JL Vigdor. "The Academic Achievement Gap in Grades 3 through 8." NBER Working Paper No. 12207 (May 2006)
- [COH92] J Cohen. "Quantitative Methods in Psychology: A Power Primer." *Psychological Bulletin*. (1992):155-159.
- [DAV94] JE Davis and WJ Jordan. "The Effects of School Context, Structure, and Experiences on African American Males in Middle and High School." *The Journal of Negro Education*, Vol. 63:4 (Autumn, 1994): 570-587.
- [ENG06] N Engec. "Relationship Between Mobility and Student Performance and Behavior." *The Journal of Educational Research*, Vol. 99:3 (Jan-Feb 2006): 167 – 178.
- [FER01] RF Ferguson. "A Diagnostic Analysis of Black-White GPA Disparities in Shaker Heights, Ohio." *Brookings Papers on Education Policy* (2001) 347-414

- [FER02] RF Ferguson. "What Doesn't Meet the Eye: Understanding and Addressing Racial Disparities in High-Achieving Suburban Schools." November 2002. Retrieved from <http://www.ncrel.org/gap/ferg/> on May 13, 2007.
- [FER98] Ferguson, R. F. (1998). "Teachers' perceptions and expectations and the black-white test score gap." In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*: 273-317. Washington, DC: Brookings Institution.
- [FRY06] RG Fryer and SD Levitt. "The Black-White Test Score Gap Through Third Grade," *American Law and Economics Review*, Oxford University Press, vol. 8:2(2006): 249-281.
- [GAR92] AM Garibaldi. Educating and Motivating African American Males to Succeed. *The Journal of Negro Education*, Vol. 61, No. 1. (Winter, 1992): 4-11.
- [GRA04] TC Grantham. "Multicultural Mentoring to Increase Black Male Representation in Gifted Programs." *The Gifted Child Quarterly*. Vol.48:3 (2004): 232:45.
- [HAN01] EA Hanushek, JF Kain, JM Markman, and SG Rivkin, "Does Peer Ability Affect Student Achievement?" *NBER Working Paper No. W8502*. (Oct 2001).
- [HAN03] EA Hanushek, "The Failure of Input-based Schooling Policies," *Economic Journal*, 113 (Feb 2003): F64-F98
- [HAN06] EA Hanushek and SG Rivkin. "School Quality and the Black-White Achievement Gap," *NBER Working Papers 12651* National Bureau of Economic Research, Inc. (2006)
- [HAR06] DN Harris and CD Herrington. "Accountability, Standards, and the Growing Achievement Gap: Lessons from the Past Half-Century." *American Journal of Education*, Vol. 112:2 (Feb 2006): 209 -238.
- [HOD73] G. Hodgson, "Do Schools Make a Difference?" *The Atlantic*, March (1973): 35-46.
- [ISA03] Illinois State Board of Education. *The Illinois State Assessment 2003 Technical Manual*. Retrieved on July 30, 2007 from http://www.isbe.state.il.us/assessment/pdfs/isat_tech_2003.pdf.
- [ISA06] Illinois State Board of Education. *Report on the ISAT/SAT-10 Bridge Study and Development of the 2006 ISAT Reporting Scales*.

Retrieved November 9, 2006 from
http://www.isbe.state.il.us/assessment/pdfs/Bridge_Study.pdf

- [JAC02] J Jacobson et al. *Educational Achievement and Black-White Inequality*. Washington, DC : National Center for Education Statistics (2002)
- [JEN98] CR Jencks and M Phillips, Eds. “*The Black-White Test Score Gap*” Washington, DC: The Brookings Institute, (1998).
- [KOB01] N Kober. *It Takes More than Testing: Closing the Achievement Gap*. Center on Education Policy (2001).
- [LIE76] S Lieberman. “Rank-Sum Comparisons between Groups” *Sociological Methodology*, vol. 7: 276-291.
- [MAN87] GK Mandeville and LW Anderson. “The Stability of School Effectiveness Indices across Grade Levels and Subject Areas.” *Journal of Educational Measurement*: Vol. 24:3.(Autumn, 1987):203-216.
- [MSA02] Minority Student Achievement Network. *Ed-Excel Assessment of Secondary School Student Culture*. (2002) Retrieved from <http://www.msanetwork.org/pub/edexcel.pdf>.
- [OGB03] JU Ogbu. *Black American Students in an Affluent Suburb: A Study of Academic Disengagement* Mahwah, NJ: Lawrence Erlbaum Associates, 2003.
- [RUM98] RW Rumberger and KA Larson. “Student Mobility and the Increased Risk of High School Dropout” *American Journal of Education*, Vol. 107:1 (Nov. 1998):1-35
- [SAN96] WL Sanders and JC Rivers. *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*. (1996) Knoxville: University of Tennessee Value-Added Research and Assessment Center
- [SCH06] DJ Schemo. “Public-School Students Score Well in Math in Large Scale Government Study”. *The New York Times*, January 28, 2006
- [SEL94] MH Seltzer, KA Frank, and AS Bryk. “The Metric Matters: The Sensitivity of Conclusions about Growth in Student Achievement to Choice of Metric.” *Educational Evaluation and Policy Analysis*., Vol. 16:1 (Spring 1994): 41-49.
- [SPE07] M Spellings. “Letter to Chief State School Officers providing an update on implementation of the No Child Left Behind Act” Dated April 23, 2007.

Retrieved July 5, 2007 from
<http://www.ed.gov/print/policy/elsec/guid/secletter/070423.html>.

- [VIA06] D Viadero. "Race Report's Influence Felt 40 Years Later: Legacy of Coleman Study Was New View of Equity." *Education Week*, Vol.25:41:1 (21-24 Jun 2006)
- [WEN02] H Wenglinsky. "The Link between Teacher Classroom Practices and Student Academic Performance." *Education Policy Analysis Archives* Vol:10 (2002):12
- [WRI97] SP Wright, SP Horn, and WL Sanders. "Teacher and classroom context effects on student achievement: Implications for teacher evaluation." *Journal of Personnel Evaluation in Education*, Vol. 11:1 (1997):5767.