THE MEASUREMENT AND DECOMPOSITION OF ACHIEVEMENT EQUITY:
AN INTRODUCTION TO ITS CONCEPTS AND METHODS INCLUDING A
MULTIYEAR EMPIRICAL STUDY OF SIXTH GRADE READING SCORES

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in the
Graduate School of Ohio State University

By
Francis H. Rogers, M.A.

*****

Ohio State University
2004

Dissertation Committee:
Professor William Loadman, Advisor
Professor Helen Marks
Professor Scott Sweetland

Approved by

Advisor
College of Education
ABSTRACT

The maintenance of democracy requires the adequate education of our youth. Americans have struggled since the founding of the republic to define the scope of this democratic imperative and to fulfill its requirements. Over the course of the twentieth century, the educational minimum was defined in terms of fiscal resources for education. Educational equality, in turn, has been measured in terms of the equality of the distribution of fiscal resources.

This paper contends that the shift away from fiscal equity toward achievement equity calls for a shift in our approach to the measurement of educational equality. The study expands the conceptual framework for measuring educational equality by incorporating the ideas of the democratic imperative for education and standards-based reform. These two ideas are brought together in the concept of egalitarian inequality, a condition marked by all students reaching or exceeding the level of educational attainment necessary to function in a democratic society.

Drawing on this framework and established measures of equality, the study offers an empirical analysis of achievement equity. Posing the question of whether achievement equity has improved in recent years, as we would expect based on state and federal accountability legislation, we examine changes in equity in the distribution of reading proficiency scores drawn from a large, Midwestern school district for the years
1998 through 2003. Characteristics of the equality measurement techniques are described and compared. Statistical significance of observed changes in equity values is tested based on observed variability in student achievement. Following the federal mandate for disaggregating student subgroup achievement, decomposition techniques are applied to determine within- and between-groups contributions to district equity. Results of decomposition measures are presented and compared.

Implications for policy makers and equity analysts are discussed with regard to gaining a more systematic understanding of the achievement gap and how, or whether, schools and districts are making progress toward closing those gaps. Recommendations for future work are offered both for broader, state and national studies and for more detailed district, school, and subgroup decompositions.
Dedicated to
George Drayton Strayer:
Father of the measurement of educational equality.
ACKNOWLEDGMENTS

Research, like all forms of learning, is an arduously challenging process, fighting outwardly at all times against a surrounding sea of ignorance while struggling inwardly against the constraints of tradition, authority, and certainty. A few elements of today’s research record will persist, much will be revealed for the nonsense that it is, and some of that will be redeemed only after rediscovery. We smile at the grave certainty of the 14th century astronomers just as others will one day smile at us. For all its difficulties and uncertainty we stay on the path of systematic study and learning in the hope of gaining - and of offering - a better understanding of ourselves, each other, and the world in which we live. As Will Durant observed in his review of the life and works of Thomas Aquinas, “The peace that surpasseth understanding is understanding” (Durant, 1950).

In offering the following manuscript to you, patient reader, I acknowledge first and foremost that I have borrowed much and contributed only a very little. I gratefully thank all those by whose efforts my own have prospered. To their great river of learning I cast these few drops.

And who are those mentors, the names of those whose lives and works have conspired with time and circumstance to enable the present study? Who were those learned Phoenician scholars who formulated the rules of the alphabet whose heritage is scattered across every page? Who were those Hindu masters whose sublime skills I have
inherited in the number system our historians mistakenly call Arabic? And who were those patient Arab doctors whose formulations led to the development of the algebra by which every analysis in this study is enabled? I acknowledge first and foremost these teachers who, though their names have been forgotten by the carelessness of time, yet breathe life on every page of this dissertation. To this legacy of influences I would add the names of:

Sister Carmel, because learning is a blessed delight;

Mr. Jack Hurd, because a second chance is a difficult thing to give;

Dr. Richard Shiels, because understanding who we have been helps us better understand who we are becoming;

Dr. Robert Klingensmith, because we don’t just live – we live someplace;

Dr. James O. Wheeler, because good work is hard work;

My mother, because ‘why?’ is a secret passage to a world of adventure;

This dissertation would not have been possible without the direct help of many very special people. I would like to acknowledge the patient support and critical review offered by my committee members, Professor Helen Marks and Professor Scott Sweetland. To my advisor and committee chairman, Professor William Loadman, I owe a particularly great debt of thanks for continually challenging my ideas and efforts. Thank you, Bill. And a very heartfelt thanks to Dr. Mary Peters who provided not only the data for the analyses but also a friend’s unflagging confidence. Thank you, Mary, for your support and your friendship.

In each of our lives we are touched by innumerable, special people whose support, guidance, humor, and wisdom help us reach out and grow. Don Williams, Mary
Mendoza, Dennis Kempel, Pete Petrlich, Pejmaan Fallah, Barbara Boggs, and Mary Jonaitis are a few of the people by whose examples I continue to study the most important problem of all, which is how to become a better person.

My sisters, Diane, Vonnie, Pat, Renee, and Elayne, having failed in their plots to assassinate me at a young age, have turned out to be okay after all. It’s funny, isn’t it, how we are close to each other at different times and in different ways. Thank you, ladies, for sparing my life, for now anyway, and thank you for your support.

Chris, my brother, for all the grief we shared and for as different as our paths have been it’s funny now how similar we really are. Not that I am ready to sell everything and start raising sled dogs with you in that frozen wasteland you call paradise, but you and I have each chosen a star to follow, and have followed where those stars have led. How many people know that kind of freedom? Stay warm.

Finally, to two of the most special people anyone could ever hope to meet: Uncle Tom and Aunt Marilyn, you will never know how much happiness you have given, not only to me but to everyone around you. Your pictures are printed all over my dictionary, next to words like courtesy, faith, honesty, integrity, kindness, and most of all, love. You have both been wonderful role models and I love you very much.

Ok, Wendy, it’s time to go camping. And yes, I’ll be doing the cooking now.
VITA

1961 ............................................................ Born – Kent, Ohio

1976 – 1982................................................. Partner
D & F Painting

1979 – 1984................................................. Operations Manager
Balloon Tours of America

Ohio State University

1985-1986 ................................................... Undergraduate Research Assistant
Department of History
Ohio State University

1987-1988 ................................................... Undergraduate Research Assistant
Department of Geography
Ohio State University

1988 – 1990................................................. Graduate Teaching Assistant
Cultural and Physical Geography
University of Georgia

1990 ............................................................ M.A. Economic Geography
University of Georgia

1991 ............................................................ Assistant Planner
Mid-Ohio Regional Planning Commission

1992 ............................................................ Associate Planner
Mid-Ohio Regional Planning Commission

1993 ............................................................ Senior Planner
Mid-Ohio Regional Planning Commission
1993 – 1994................................................. Associate Policy Analyst
Ohio Department of Education

1994 – 1996................................................. Senior Policy Analyst
Ohio Department of Education

1996 – 2001................................................. Assistant Director, Policy Research
Ohio Department of Education

2001 – 2002................................................. University Fellow
Ohio State University

2002 ............................................................ M.A. Quantitative Research Methods
Ohio State University

2002 – 2003................................................. Graduate Teaching Assistant
Experimental Design
Ohio State University

2003 – 2004................................................. Graduate Research Assistant
Ohio Partnership for Accountability
Ohio State University

2001 – present ............................................. Principal
The Policy Research Group

MAJOR FIELDS OF STUDY

Major Field:  Education
Educational Statistics
Measurement
Evaluation

Cognate:    Policy Analysis
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract ....................................................................................................................... ii</td>
</tr>
<tr>
<td>Dedication ..................................................................................................................... iv</td>
</tr>
<tr>
<td>Acknowledgments ........................................................................................................... v</td>
</tr>
<tr>
<td>Vita ........................................................................................................................... viii</td>
</tr>
<tr>
<td>List of Tables ................................................................................................................. xv</td>
</tr>
<tr>
<td>List of Figures .............................................................................................................xvi</td>
</tr>
</tbody>
</table>

Chapters:

1. Introduction ............................................................................................................... 1
   
   1.1 Overview ........................................................................................................ 1
   1.2 Background: Trends in educational equality .................................................. 2
   1.3 Educational adequacy .................................................................................... 4
   1.4 The achievement gap ..................................................................................... 6
   1.5 Can equity be measured? .............................................................................. 7
   1.6 Is educational achievement becoming more equal? ...................................... 9
   1.7 Statement of problem .................................................................................... 11
   1.8 Conceptual framework ............................................................................... 12
   1.9 Purpose and objectives of the study ............................................................ 12
   1.10 Limitations .................................................................................................. 14
   1.11 Assumptions .............................................................................................. 16
   1.12 Contributions to the field ........................................................................... 17
   1.13 Layout of the study ................................................................................... 18

2. Review of literature .................................................................................................... 20
   
   2.1 The democratic imperative for education ...................................................... 20
   2.2 The struggle for educational equity ............................................................... 23
   2.3 Achievement equalization ............................................................................. 27
      2.3.1 Testing and educational achievement .................................................... 28
      2.3.2 *A Nation At Risk* and the decade of systemic reform: 1983 – 1993 .......... 31
      2.3.3 The nationalization of standards-based reform: 1994 – 2002 ................. 32
2.3.4 No Child Left Behind Act of 2001: National mandate for achievement equity ..........................................................33
2.4 The measurement of educational equality ..........................................................35
  2.4.1 Current practice: School report cards as measures of achievement equality .................................................36
  2.4.2 The measurement of income equality ..............................................................................................................38
  2.4.3 The measurement of educational equity: From Cubberly to Berne & Steifel .................................................44
  2.4.4 Within- and between-groups measures of equality ..........................................................................................52
  2.4.5 The measurement of achievement equity .........................................................................................................53
2.5 Developing an achievement equity research agenda ...........................................54
  2.5.1 Proposal for empirical study .........................................................................................................................56
2.6 Chapter summary ..................................................................................................57

3. Methodology ...........................................................................................................60
  3.1 Introduction ............................................................................................................60
  3.2 Study period .........................................................................................................61
  3.3 Geographical coverage .......................................................................................61
  3.4 Student grouping variable ..................................................................................62
  3.5 Student achievement measure ............................................................................62
  3.5.1 Ohio proficiency tests .....................................................................................62
  3.5.2 Reliability .........................................................................................................63
  3.5.3 Validity ..............................................................................................................65
  3.5.4 Reliability of the Ohio 6th grade test of reading proficiency .................................................66
  3.5.5 Validity of the Ohio 6th grade test of reading proficiency .................................................66
  3.6 Population ..........................................................................................................70
  3.7 Source of data .....................................................................................................71
  3.8 Missing data .........................................................................................................71
  3.9 Research questions ...............................................................................................72
  3.10 Analytical method ...............................................................................................72
  3.11 Evaluation of statistical hypotheses ......................................................................74
  3.11.1 Research Question 1 ......................................................................................74
  3.11.2 Research Question 2 ......................................................................................77
  3.11.3 Research Question 3 ......................................................................................80
  3.12 Computational characteristics of the equity measures .......................................80
  3.12.1 Point measures of horizontal equity ................................................................81
  3.12.1.1 Range .........................................................................................................81
  3.12.1.2 Restricted range ........................................................................................82
  3.12.1.3 Federal range ratio ....................................................................................83
  3.12.1.4 Summary of the point measures of horizontal equity ........................................84
  3.12.2 Variance-based measures of horizontal equity ................................................84
4. Research findings ................................................................................................. 104

4.1 Introduction ........................................................................................................ 104

4.2 Research Question 1: Has performance increased in the district overall? ................................................................. 105

4.2.1 Decomposable measures of horizontal equity ........................................ 105

4.2.2 Means and standard deviations ................................................................. 107

4.2.3 Analysis of variance .................................................................................... 109

4.2.3.1 Assumption checks for ANOVA ...................................................... 111

4.2.3.2 Dependent variable measured at interval or ratio level .................. 111

4.2.3.3 Independent observations ............................................................... 112

4.2.3.4 Randomization ............................................................................... 114

4.2.3.5 Normality of underlying population distributions ........................... 115

4.2.3.6 Equality of underlying population variances .................................. 116

4.2.3.7 Summary of assumption checks .................................................... 117

4.2.3.8 Omnibus test of equality of means ............................................... 117

4.2.3.9 Pairwise post hoc comparisons to base year .................................. 119

4.2.3.10 Post hoc test of linear trend ......................................................... 121

4.2.4 Summary of evidence regarding Research Question 1 ....................... 123

4.3 Research Question 2: Have reading scores converged across the district? ........................................................................... 123

4.3.1 Descriptive measures of horizontal equity ............................................ 124

4.3.1.1 Range .......................................................................................... 125

4.3.1.2 Restricted range ........................................................................... 125
4.3.1.3 Federal range ratio ................................................................. 126
4.3.1.4 Summary of the descriptive measures of horizontal equity ............................................................ 127
4.3.2 Variance-based measures of horizontal equity ............................................................ 128
4.3.2.1 Standard deviation ................................................................. 128
4.3.2.2 Coefficient of variation .......................................................... 129
4.3.2.3 Standard deviation of logarithms ............................................ 130
4.3.2.4 Summary of variance-based measures of horizontal equity ............................................................ 132
4.3.3 Distributional measures of horizontal equity ............................................................ 132
4.3.3.1 Quintile analysis ................................................................. 132
4.3.3.2 Gini coefficient .............................................................. 138
4.3.3.3 Atkinson index .............................................................. 140
4.3.3.4 Theil statistic ................................................................. 145
4.3.3.5 Summary of the distributional measures of horizontal equity ............................................................ 147
4.4 Research Question 3: Have reading scores converged within and between schools? ............................................................ 149
4.4.1 Descriptive statistics by school ................................................... 149
4.4.2 Atkinson decomposition .............................................................. 159
4.4.3 Theil decomposition .................................................................... 162
4.4.4 Summary of the within- and between-schools equity decompositions ............................................................ 164
5. Review of findings ........................................................................................... 165
5.1 Introduction ............................................................................................... 165
5.2 Summary of findings ........................................................................................... 165
5.2.1 Research Question 1 ............................................................................... 166
5.2.1.1 Pass rates and descriptive statistics .............................................. 166
5.2.1.2 Assumption checks for ANOVA .................................................. 167
5.2.1.3 Results of ANOVA ................................................................. 168
5.2.1.4 Summary of results for Research Question 1 .................................. 169
5.2.2 Research Question 2 ............................................................................... 170
5.2.2.1 Descriptive statistics .................................................................. 170
5.2.2.2 Point measures of horizontal equity ............................................... 170
5.2.2.3 Variance-based measures of horizontal equity .................................................. 171
5.2.2.4 Distributional measures of horizontal equity .................................................. 171
5.2.2.5 Summary of results for Research Question 2 .................................. 173
5.2.3 Research Question 3 ............................................................................... 173
5.2.3.1 Descriptive statistics .................................................................. 174
5.2.3.2 Summary of results for Research Question 3 .................................. 174
6. Conclusions and recommendations

6.1 Introduction

6.2 Discussion
   6.2.1 The equity measures
   6.2.2 Decompositions
   6.2.3 The study district
   6.2.4 The instrument
   6.2.5 Generalizability

6.3 Limitations

6.4 Theoretical, practical, and research implications
   6.4.1 Theoretical implications
   6.4.2 Practical implications
   6.4.3 Research implications

6.5 Reflections

Bibliography

Appendix A: Raw score to scaled score conversion table, grade 6 reading,
March 2003
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Statewide average, reliability, and standard error of measure for the 6th grade reading test</td>
</tr>
<tr>
<td>4.1</td>
<td>Sixth grade reading proficiency rates, 1998-2003</td>
</tr>
<tr>
<td>4.2</td>
<td>Descriptive statistics for 6th grade reading, 1998-2003</td>
</tr>
<tr>
<td>4.3</td>
<td>Shapiro-Wilk test of normality of reading scale scores by year, 1998-2003</td>
</tr>
<tr>
<td>4.4</td>
<td>Skewness statistics for reading scale scores, 1998-2003</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary of assumption checks</td>
</tr>
<tr>
<td>4.6</td>
<td>Summary of one-way ANOVA of reading scale scores by year</td>
</tr>
<tr>
<td>4.7</td>
<td>Summary of Dunnett test of reading scale score means, 1998 vs. 1999-2003</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary of Scheffè pairwise test of reading scale score means, 1998-2003</td>
</tr>
<tr>
<td>4.9</td>
<td>Summary of one-way ANOVA linear trend analysis of reading scale scores across years, 1998-2003</td>
</tr>
<tr>
<td>4.10</td>
<td>Descriptive measures of equity for 6th grade reading, 1998-2003</td>
</tr>
<tr>
<td>4.11</td>
<td>Summary of descriptive measures of horizontal equity for 6th grade reading, 1998-2003</td>
</tr>
</tbody>
</table>
4.14 Quintile means for reading scale score, 1998-2003
4.15 Annual proportions of scores by quintile for 6th grade reading, 1998-2003
4.16 Gini coefficient for 6th grade reading scale scores, 1998-2003
4.17 Atkinson index for 6th grade reading scale scores, 1998-2003
4.18 Atkinson index of 6th grade reading scale scores, 1998-2003 for e=30
4.19 Theil statistic for 6th grade reading scale scores, 1998-2003
4.20 Summary of distributional measures of horizontal equity for 6th grade reading scale scores, 1998-2003
4.21 Descriptive statistics for 6th grade reading scale scores by school, 1998-2003
4.22 Summary of one-way ANOVA of reading scale scores by school, 1998-2003
4.23 Atkinson decomposition of within- and between-schools equity, 1998-2003
4.24 Theil decomposition of within- and between-schools equity of 6th grade reading scale scores, 1998-2003
A.1 Raw score to scaled score conversion table, grade 6 reading, March, 2003
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Lorenz curve</td>
<td>93</td>
</tr>
<tr>
<td>4.1</td>
<td>Plot of 6th grade reading proficiency rates, 1998-2003</td>
<td>107</td>
</tr>
<tr>
<td>4.2</td>
<td>Plot of 6th grade reading scale score means by year, 1998-2003</td>
<td>109</td>
</tr>
<tr>
<td>4.3</td>
<td>Standard deviation of reading scale scores, 1998-2003</td>
<td>129</td>
</tr>
<tr>
<td>4.4</td>
<td>Coefficient of variation for reading scale scores, 1998-2003</td>
<td>130</td>
</tr>
<tr>
<td>4.5</td>
<td>Standard deviation of logarithms for reading scale scores, 1998-2003</td>
<td>131</td>
</tr>
<tr>
<td>4.6</td>
<td>Quintile means for reading scale scores, 1998-2003</td>
<td>134</td>
</tr>
<tr>
<td>4.7</td>
<td>Relative proportion of 6th grade reading scale scores by quintile, 1998-2003</td>
<td>137</td>
</tr>
<tr>
<td>4.8</td>
<td>Observed and benchmark Gini coefficient for 6th grade reading scale scores, 1998-2003</td>
<td>139</td>
</tr>
<tr>
<td>4.9</td>
<td>Atkinson index for 6th grade reading scale scores, 1998-2003</td>
<td>142</td>
</tr>
<tr>
<td>4.10</td>
<td>Atkinson index for 6th grade reading scale scores, 1998-2003 for e=30</td>
<td>144</td>
</tr>
<tr>
<td>4.11</td>
<td>Theil statistic for 6th grade reading scale scores, 1998-2003</td>
<td>146</td>
</tr>
<tr>
<td>4.12</td>
<td>Mean reading scale score for School 1, 1998-2003</td>
<td>153</td>
</tr>
<tr>
<td>4.13</td>
<td>Mean reading scale score for School 2, 1998-2003</td>
<td>153</td>
</tr>
<tr>
<td>4.14</td>
<td>Mean reading scale score for School 3, 1998-2003</td>
<td>154</td>
</tr>
</tbody>
</table>
4.15 Mean reading scale score for School 4, 1998-2003................................. 154
4.16 Mean reading scale score for School 5, 1998-2003................................. 155
4.17 Mean reading scale score for School 6, 1998-2003................................. 155
4.18 Mean reading scale score for School 7, 1998-2003................................. 156
4.19 Mean reading scale score for School 8, 1998-2003................................. 156
4.20 Mean reading scale score for School 9, 1998-2003................................. 157
4.21 Mean reading scale score for School 10, 1998-2003............................... 157
4.22 Mean reading scale score for School 11, 1998-2003............................... 158
4.23 Mean reading scale score for School 12, 1998-2003............................... 158
4.24 Atkinson decomposition of within- and between-schools equity of 6th grade reading scale scores, 1998-2003......................................................... 161
4.25 Theil decomposition of within- and between-schools equity for 6th grade reading scale scores, 1998-2003.......................................................... 163
CHAPTER 1

INTRODUCTION

1.1 Overview

This paper addresses a basic but crucial question regarding the measurement of change among student achievement scores over time. The paper offers a set of methods for understanding, for a group or set of groups of students, e.g., schools, districts, subgroups of students, 1) whether student scores are rising, and 2) whether the distribution of scores is shrinking, particularly as a result of gains among lower-performing students. These two questions lie at the heart of our current achievement accountability policies. As we demonstrate below, however, our current measures have critical limitations that certain well-established equity measures directly inform. The paper approaches the problem from the perspective of educational equity, tracing the democratic imperative for educational equality from the founding of the republic through the adoption of the *No Child Left Behind Act of 2001* (NCLB). After examining and critiquing the current methods of measuring progress in achievement, this study offers an alternative approach that promises to provide more comprehensive information regarding the sought-for changes in student achievement.
1.2 Background: Trends in educational equality

How equally are students sharing the benefits of the public investment in K-12 education? This question, in various ways at different times, has concerned policy makers and educators since the founding of the republic. Thomas Paine, Thomas Jefferson, Noah Webster, and Benjamin Rush, drawing on such diverse philosophical perspectives as those of John Locke in his Second Treatise of Government (1679, reprinted and discussed in Thomas, 1995) and Adam Smith’s Wealth of Nations (1776), agreed in spite of wide differences on a multitude of philosophical and political questions on the imperative of a general system of education to ensure the adequate education of the electorate as the key means of safeguarding and nurturing America’s democratic institutions and systems of self government. By the 1820’s, James G. Carter, America’s ‘Father of Normal Schools’ (Hinsdale, 1900, p. 55) was decrying wealth-based disparities in public education (Carter, 1826), and Horace Mann was taking up the cause of educational equality through his leadership of the common school movement, directed in part to the standardization of educational opportunity. Efforts to build a system of public primary and secondary education continued through the nineteenth century, with state constitutional conventions and legislatures adopting educational mandates for ‘adequate’ ‘thorough’ and ‘efficient’ education. In spite of these efforts, concerns for educational equality persisted.

Severe disparities in educational equality, particularly in the South led eventually to the watershed case of Brown v. Board of Education (1954) in which the United States Supreme Court finally intervened, overturning racial segregation of schools and mandating equal educational opportunity for all. Through this action the issue of
educational equity emerged as an urgent matter of public concern. Invoking the equal opportunity clause of the 14th Amendment, *Brown* opened the door not only to challenges to race-based disparities, but to disparities among other legal classes. The scope of the problem expanded to include a broad array of interests as reflected in a long series of federal and state litigation spanning problems of discrimination and educational inequality based on sex, disability, economic disadvantage, and property wealth. Equity litigation centered on equality of access and equality of educational resources, from entry to the schoolhouse door to the condition of the schoolhouse itself, from appropriate services to equality in the general level of fiscal resources across districts.

By the 1990’s, the struggle for equality in education had begun shifting away from the problem of equalizing educational inputs. Standards-based reform in education, beginning in the 1960’s with the evaluation requirements embedded in the *Elementary and Secondary Education Act of 1965* and the adoption in Florida in 1969 of a statewide system of educational accountability built around minimum competency testing, had expanded rapidly in the 1970’s and 1980’s. By 1990 the majority of states had adopted standards based assessment and accountability systems, a criterion referenced approach to the evaluation of student acquisition of state-defined academic content. States moved in different directions, but the general trend by the beginning of the 1990’s was toward state-enforced, district accountability, whereby districts were required to meet state-defined student achievement targets, typically passage rates on criterion referenced tests, or face state sanctions that might go as far as state takeover or reorganization of the local school district. By the end of the 1990’s, states were expanding accountability from the district to the school level. The federal government began taking a more active role as
the Clinton administration sought to tie school accountability to federal funding under the Elementary and Secondary Education Act (ESEA). The federalization efforts, begun by Clinton, were continued by the Bush administration, culminating in 2002 with the signing into law of the ESEA reauthorization under the name of the No Child Left Behind Act of 2001. This bipartisan legislation established a permanent federal role in performance accountability, mandating that all states adopt performance standards and that 95% of all students meet those standards by the year 2014.

From the perspective of educational equity, the accountability movement integrated the multiple, separate struggles for educational equalization. With requirements for ensuring adequate achievement across all students, by race, sex, ability, economic status, and English proficiency, standards based accountability addressed the separate concerns of all of these groups. State adopted standards represent a democratically determined set of educational outputs defining the minimum level of educational attainment needed by a child in order to undertake successfully the responsibilities of democratic citizenship and to become a productive community member.

1.3 Educational adequacy

Paralleling the expansion of the standards-based reform movement and the trend toward state systems of district and school accountability, the trend in equity litigation and adjudication began in 1990 to shift away from the concern with the overall distribution of inputs across the educational system, focusing instead on the issue of educational adequacy (Kentucky v. Rose, 1990). What this meant was that jurists and
litigants turned from using educational resources as the key measure of educational opportunity to using student acquisition of a defined set of knowledge, skills and abilities. Educational inputs were not ignored altogether, however, but were simply recast. The problem of whether all students were equally situated with respect to the educational fiscal investments was transformed to the problem of whether or not all students received at least a minimally adequate share of those resources sufficient to acquire the basic skills, knowledge and abilities needed in order to become effective participants in democratic society and productive members in the free market (Koski & Levin, 2000).

In effect, the standards reform movement operationalized in each state a set of educational outputs on which the courts and litigants would come to rely in the legal contests concerning educational equity and the state’s role in assuring adequate educational opportunity to public school students. Using a state’s performance measures, plaintiffs demonstrated, using a set of commonly administered assessments plus commonly measured attendance and graduation rates, a correlational relationship between local fiscal resources and local student performance. Plaintiffs have used this information as evidence of a causal link between resources and performance and, as a consequence, that funding disparities result in educational achievement disparities. Subsequently, plaintiffs, pointing to performance in higher-expenditure school districts, have argued for a link between funding level and the level of performance desired from the system. In Ohio, for example, the state supreme court in DeRolph v Zelman (2000) accepted a definition of resource adequacy derived from the average expenditures of school districts meeting the set of performance accountability standards mandated under Ohio law.
1.4 The achievement gap

The implementation of state systems of educational performance accountability and the concern for educational equality revealed wide disparities in student performance both across school districts and across certain subgroups of students (Ferguson, 1998a; 1998b). Variation in student scores is not, of itself, problematic. The standards are egalitarian only to the limited extent that students are expected to meet a minimum threshold of proficiency; the standards ensure the opportunity for access to the democratic process while recognizing students individual differences (Britell, 1980). The standards accommodate inequality to the degree that the objective is for students to meet or exceed the standards; any range of variation might exist above the level of the minimum standards without implying concern for the egalitarian standard. The standards may be understood therefore to reflect a goal of egalitarian inequality in student achievement where variation is acceptable so long as it is above the test’s minimum cut score.

Of concern to policymakers and educators has been the fact that certain patterns of sub-standard performance emerged in the test data (AAUW, 1992; Howard, 2003; Hedges & Nowell, 1998). Geographically, sub-standard achievement tends to be concentrated in urban and poor rural districts. Demographically, sub-standard performance tends to be higher among Hispanics than whites and higher still among blacks. Although test results have been associated with calls for educational reform since Mann’s first efforts to link the two in the early 19th century, the publication of results under the state’s standards-based accountability systems have led to calls not just from within education but from the state and federal policy arenas for educational reforms
ranging from curriculum redesign to community network building, teacher preparation program evaluation, adoption of quality-improvement principles, national standards for teacher licensure and cultivation of increased parental participation, to name a few.

1.5 Can equality be measured?

The measurement of equality can be traced to studies of income inequality beginning as early as the 1880’s, with economists devising techniques by which to judge the comparative equality of two or more distributions of income data (e.g., Goschen, 1887; Atkinson, 1888; Wolf, 1892; Gini; 1912; Dalton, 1920; Theil, 1967; Atkinson, 1970; Sen, 1997; Conceição & Ferreira, 2000; de la Vega & Urrutia, 2003). The problem appeared simple at first. Measures such as the standard deviation (Goschen, 1887) or the interquartile range (Bowley, 1902), for example, suggest whether two distributions are equally concentrated or dispersed. As conceptual development continued, with researchers gaining access to a growing body of empirical examples, it became clear that different measures capture different characteristics of the equality of, or differences between distributions. In 1920 Baron Hugh Dalton offered a general conceptual framework and typology of measures for the measurement of income equality (Dalton, 1920). Since that time, researchers have continued to refine the methods and concepts of equality measurement (e.g., Theil, 1967; Atkinson, 1970; Sen, 1997; Conceição & Ferreira, 2000; Lasso de la Vega & Urrutia, 2003).

Efforts to quantify equality in education emerged in the early 1900’s in response to growing concerns regarding the fairness in the distribution of educational resources, achievement, and post-secondary opportunities (Cubberly, 1905; Strayer, 1905). The
initial focus was on the equitable distribution of educational resources, and of fiscal resources in particular. Beginning with Ellwood Cubberly’s (1905) enumeration of public school inequalities and George Strayer’s (1905) pioneering work in the measurement of educational equity, researchers focused on the concern that America’s system of primary and secondary education was resulting in widely disparate levels of educational service, achievement, and opportunity determined by the fortunes of an administrative geography of what had already by that time become the largely autonomous public school districts. Although policy makers attempted to provide relief to the poorest school districts through the adoption of state funded equalization programs (Cubberly, 1905; Strayer, 1905; Strayer & Haig, 1923; Strayer, Haig, & Mort, 1924), wide disparities continued.

In the 1960’s and 1970’s educational finance experts began to develop increasingly systematic equality measurement methodologies (Harrison & McLoone, 1965; Hickrod & Sabulao, 1969; Mushkin & McLoone, 1965; McLoone, 1974). The long series of legal challenges to inequality in state funding for education created demand for quantitative measures of equality by which plaintiffs, defendants, and the courts could systematically distinguish the degree of equality achieved in a given state education system. Drawing on the rich body of methodological literature developed by economists interested in problems of income inequality (a summary to that period is provided by Sen, 1973), educational economists adapted measures of distributional equality to the measurement of equality in educational finance data (Odden, Berne, & Steifel, 1979; Berne & Steifel, 1984; Odden & Picus, 2000). Consistent with the primary focus on the distribution of fiscal resources for education, the equity measures were employed in the
analysis of educational revenues and expenditures, collectively described as educational ‘inputs’ (Odden, Berne, & Steifel, 1979).

1.6 Is educational achievement becoming more equal?

Given the length of time states have had to raise student performance in accordance with state-defined, locally implemented content frameworks and assessments, is there evidence that achievement has indeed become more equal over time? Educational researchers have demonstrated that the techniques for measuring income equality are useful in the analysis of distributions of educational data, with an extensive literature having emerged on problems of educational fiscal equality (Berne & Stiefel, 1984; Odden & Picus, 2000). While the emphasis in educational research has been the distribution of fiscal resources, referred to in that literature as educational inputs, researchers have acknowledged that, conceptually at least, the methods may logically be extended to the evaluation of distributions of achievement data, referred to generally as educational outputs (Berne & Stiefel, 1984).

The current method of measuring achievement and change in achievement over time is the student pass rate on state exams (Ohio Department of Education, 2004). Increasing pass rates are assumed to represent forward progress of the general student body while convergence among pass rates is assumed to represent a decrease in achievement differentials between student subgroups. The pass rate describes conditions at a single point in a distribution of student scores: the proportion of students above and below that point. The pass rate does not describe the spread of the distribution, i.e., the similarity of student scores, nor does the pass rate describe whether, over time, the degree
of similarity among scores is increasing or decreasing. A simple example will illustrate this problem.

A hypothetical school district has 100 fourth graders in each of two years. Three fourths of the students in each year are white, one quarter are black. The pass rate on the state’s proficiency exam in year one is 70% for whites and 50% for blacks. In year two, these rates increase to 72% for whites, 55% for blacks. Educators and policy makers applaud the results, declaring that achievement has increased and the achievement gap has declined. On closer examination, let us say that the bottom quintile of white students improved while the bottom quintile of black students declined. Did achievement really go up? Is the gap between whites and blacks actually smaller? The pass rates in this case increased because a larger proportion of students scored at or above the cut score, but the pass rates mask the fact that, for the bottom twenty percent of students at least, scores actually declined. The problem is that the pass rate is too limited a descriptive measure of student achievement to provide adequate information to answer these questions.

The equality measures from the economics literature, already demonstrated to be well suited to the analysis of fiscal equality, may also be adapted to the questions posed in the hypothetical scenario, above, regarding whether a set of distributions of student achievement data have increased and whether those distributions have converged. Of additional interest, given the accountability mandate’s application not merely to districts, but to schools within districts, and to student subgroups (race, sex, ability, economic status, and English proficiency) within schools, is a body of recent research in the decomposition of equality measures (Conceição & Ferreira, 2000; Lasso de la Vega & Urrutia, 2003), a line of inquiry that may substantially further enhance the applicability of
equality measurement to the analysis of within- and between-groups contributions to overall achievement equality.

1.7 Statement of Problem

We have discussed two questions. “How equally are students sharing the benefits of the public investment in K-12 education?” and, “Can equality be measured?” The first question is driven by the democratic imperative for education: all citizens must receive at least that level of education sufficient to prepare them for participation in democratic society. The educational equity movement responded to this problem by equalizing educational access and resources for various student groups. The standards-based reform movement, focusing on equality of educational achievement, responded to the problem of democratic education by focusing on the equalization of student achievement, or at least equalization up to the level of a publicly defined democratic minimum.

Turning to the second question, we have drawn on the work of economists and educational researchers to demonstrate that an array of measurement techniques is available. So far, however, the measurement of educational equity has been concerned with the distribution of educational inputs, particularly educational revenues and expenditures. In the face of the shift toward an achievement-based definition of educational equality, evident in public policy ranging from the standards-based reform movement and federal accountability legislation to policy responses to school finance litigation, educational researchers’ and state and federal policy makers’ knowledge of achievement equality remains limited to information provided on state report cards.
Given the intrinsic limitations of such information, this paper offers a conceptual and empirical analysis of educational equality based on student achievement scores.

1.8 Conceptual framework

In light of the articulation of state-level accountability standards across the nation, this paper offers an empirical application of equality measurement techniques to the distribution of student achievement. A standards based accountability framework implies that a responsive educational system exhibit declining variation in student performance over time by virtue of raising all students to at least the cut score on a given criterion referenced assessment of student achievement. If such an improvement process were taking place within an educational system, student scores should exhibit two measurable phenomena: a positive trend in average scores and a positive trend in equality. Extending this conceptual framework to the subgroup level, given an accountability framework that calls for improvement across disaggregated groups of students, student scores within subgroups should exhibit positive trends both in within- and between-groups equality.

1.9 Purpose and objectives of the study

Within the conceptual framework of standards based accountability this dissertation addresses three related research questions, all of which fit together under the broader question of whether or to what degree we can adapt current equality measures to the measurement of equality in the distribution of student achievement. The study provides an empirical analysis of achievement equality using sixth grade reading scores collected each year from 1998 through 2003 from a large, Midwestern school district.
The study first addresses the question of whether student scores exhibit a trend in improvement over the period. Given the weaknesses of student pass rate as an indicator of improvement over time, the study examines the utility of the analysis of trend based on the analysis of variance (ANOVA). The study thus offers a direct comparison of the pass rate and ANOVA based on an empirical example.

Secondly, the study offers an empirical analysis of achievement equity, using the same data set and a diverse set of equity measures commonly used by economists and educational finance experts. The equity measures are grouped by functional characteristic into point measures, variance-based measures, and distributional measures. A side-by-side review of results compares the consistency and distinctive characteristics of the measures.

The study proposes that an important question in the analysis of achievement equity is whether observed changes in equity over time are either statistically or substantively significant. Combining observed equity statistics and the observed variance in scores for each year, the study proposes a method of evaluating the presence of trends in equity. Using the proficiency test cut score as a benchmark reference, the study proposes a method of evaluating the degree to which observed trends are substantive.

The third objective of the study is to determine whether the analysis of achievement equity can be disaggregated to relevant subgroups. Employing decomposition techniques for the two distributional measures of equality, the study examines the level of within- and between-schools equality of reading achievement for each of the six years, 1998-2003. This empirical analysis enables evaluation of the applicability of these measures to the problem of achievement equity, how the results of
the different measures compare, and the potential feasibility of applying the measures to other grouping criteria, e.g., race, sex, ability, economic status, and English proficiency.

1.10 Limitations

As the first formal study of achievement equity, the scope of this study is necessarily limited. The study focuses on achievement equity at the district level in terms of overall district achievement and at the school level by means of the decomposition of within- and between-schools components. This design suggests numerous extensions of this work. For example, the analysis of achievement equity can be expanded up from the district level to regional, national, and even international levels (e.g., by means of such international data sets as the Third International Math and Science Study). The focus here on reading achievement can be extended to analyses both of other cognitive domains, e.g., mathematics and science, and to multivariate measures of achievement, by means possibly of discriminant analysis or factor analysis. The scope could be expanded outward to comparisons among districts, and downward to other student subgroups within districts or even to comparisons at the classroom or program level within and across schools.

This study is restricted in terms of techniques to the adaptation and application of current measures of equality. Conceivably, the study of achievement equality will reveal the need and opportunities for new measures as yet unconceived in the equality literature. Indeed, this study will conclude with several specific suggestions for additional work in the direction of adapting current techniques to more fully integrate the criterion-referenced rules governing achievement policy.
Importantly, this is not a study of causality. The interest here is in developing tools for the measurement of equality in the distribution of achievement scores rather than the analysis of causes and effects either of the underlying reading scores or of the observed equality statistics. By providing a set of tools for measuring equality in achievement, however, this study helps open the way to the study of the causality of output equity to the extent that subsequent researchers can devise and mount experimental or quasi-experimental designs that can isolate the effects of, for example, specific educational reforms or particular fiscal mechanisms on achievement equity.

The achievement measures of interest in educational accountability systems, as defined and adopted by the federal and state governments, are state-designed, criterion-referenced tests of student achievement. Such instruments offer qualified measures of student achievement, introducing problems of measurement error, reliability, and validity. Reliability, or the degree to which an instrument consistently measures the same phenomenon, can be controlled through careful construction of items in order to ensure that each item functions in a consistent and appropriate manner. Validity, or the degree to which the scores of an instrument can be interpreted in a particular way, can be controlled through careful specification of the construct of interest, e.g., sixth grade reading proficiency, selection of items that logically relate to the construct, and evidence in the form of validation studies demonstrating that the items function together in a manner consistent with the construct and any subconstructs. Chapter three addresses this issue in detail with regard to the specific measure used in this analysis, the grade six Ohio proficiency test of reading.
An important final consideration is generalizability; while the conceptual designs of the output measures might be taken to apply to any case where the assumptions and technical requirements of the measure are met, the empirical results of this study, drawn as they are from a particular district and from a particular span of time, can be assumed to be idiosyncratic unless and until further corroborating research is conducted across more districts and other states and across greater spans of time.

1.11 Assumptions

The study assumes that the achievement data used in the study are valid and reliable measures of reading proficiency. This study assumes that the Ohio proficiency tests were administered in accordance with the directions provided by the state and as mandated by law (General Assembly of the State of Ohio, Am. Sub. H.B. 182) and that the results of the test for each student in each school reflects performance under generally similar circumstances, without coaching to enhance performance and without systematic distractions or other interference to detract from performance. Implications of this assumption are discussed in detail in chapter three. Evidence relating to the assumption is presented in chapter four.

For security and validity, the state does not administer the same form of the test each year, but implements a different version each year. In using results of the sixth-grade reading test for six different years, the study assumes that the versions of the test for those six years are equivalent. Violations of this assumption would mean that differences in achievement distributions do not reflect differences in the underlying proficiency of the students but rather that the differences are merely artifacts of the
measurement instrument. This assumption is discussed in chapter three and possible violations are highlighted in chapter four.

The study assumes that the district is an appropriate unit of analysis for examining and understanding achievement equality. Since the district is chosen as illustrative for the application of the equality measurement techniques, and because the techniques are shown to be readily applicable to other units of analysis, e.g., school as the unit of analysis, this assumption is not critical to the study if it is understood as illustrative rather than normative.

1.12 Contributions to the field

This paper extends the research literature by applying the equality measurement tools to educational achievement data. The study presents those measures as a set of alternative strategies for quantifying annual student progress on state proficiency tests and convergence among subgroups on those tests. The results are shown to provide important information missing from pass rate data, information that should prove useful to educators and policymakers seeking to understand the nature and degree of change in the distribution of student achievement scores over time for cross sections of student groups and subgroups, particularly in terms of raising student achievement and reducing between-group differences.

This study provides a set of techniques for the systematic evaluation of equality in the distribution of student achievement, as measured by criterion referenced test scores, and of changes in achievement distributions across groups, places, and time.
By permitting the examination of distributions of student achievement, this study advances the analysis of educational equity from a longstanding emphasis on distributions of educational resources, which are usually taken as surrogate measures of educational opportunity, to a direct examination of achievement itself, which is the operationalization of educational opportunity that guides current public policy for educational accountability.

By providing quantitative measures of achievement equality, this study will enable future inquiry into the causal relationships between changes in achievement equality and such factors as accountability policy, school funding designs, or specific instructional programming.

From a public policy perspective, the measurement of changes in achievement equity informs the understanding of the achievement gap and the amount of progress being made to close that gap; this information can in turn inform state and federal policy responses to the equality characteristics of student achievement.

1.13 Layout of the study

The first chapter has provided the reader with an overview of the research problem and the context surrounding that problem. The second chapter explores in detail the background for the study, i.e., the origins and definition of the democratic imperative for education in America and the principle of egalitarian inequality underlying that imperative, the history of our multiple struggles to fulfill that democratic imperative and the focus in those struggles on equality in access to educational resources, a review of the techniques that have been applied to the quantification of equality of educational
resources, the shift in the struggle for educational opportunity away from equality and
toward adequacy (including standards based school reform) and a reorientation of the
equality debate toward educational achievement, and finally the emergence of the modern
educational accountability movement and the public policy frameworks designed, at least
in name, to mandate a set of democratically egalitarian standards for educational
achievement for all students. The chapter highlights two of the key policy questions
regarding the distribution of student achievement and the inadequacy of currently
reported descriptive statistics to answer those questions. The chapter concludes that what
is needed is a systematic strategy, employing appropriate equality-measurement tools, for
evaluating the distribution of student achievement in ways that address policy makers’
needs for information and which inform educators efforts to raise student achievement
across relevant subgroups.

The third chapter presents the analytical methodology, the mathematical
definitions and technical assumptions of the measures to be used, variable selection, and
data collection. Chapter four presents the results of the analysis, including a discussion
of findings and results. Chapter five provides a summary of the study results, both with
respect to the research questions and each of the measures. Chapter six provides
discussion of the study findings, conclusions drawn from the results, and
recommendations for future research.
2.1 The democratic imperative for education

The association between education and citizenship is deeply rooted in American political, legal, and education history. The Supreme Court has recognized that, ‘[Education] is the very foundation of good citizenship’ (*Brown v Board of Education of Topeka, Kansas*, 1954). Since its adoption of minimum competency achievement testing in 1977, the North Carolina legislature has continued to maintain that the state’s public and private schools must assure that high school graduates, “…possess those skills and that knowledge necessary to function independently and successfully in assuming the responsibilities of citizenship” (Section 115C-174.11(b)(1)).

We hear in these quotes the echoes of a proposition of a democratic imperative for education that can be traced, at least implicitly, to the philosophical writing of John Locke, whose words and ideas permeate a broad political and philosophical spectrum of early American republicanism and whose legacy of natural law includes both the Declaration of Independence and the Bill of Rights (see, e.g., Thomas, 1995). Flowing from the principles of Lockean liberalism, many of America’s leading founders articulated a democratic imperative for education, making explicit the tie between education and democratic citizenship (e.g., Cremin, 1980). George Washington, stepping
down from his term as President spoke to the critical need for public support of education, noting that, “In proportion as the structure of a government gives force to public opinion, it is essential that public opinion should be enlightened” (1796, quoted in Fitzpatrick, 1944, vol. 35, pp. 229-230). Such appeals for education as the cornerstone of the republic can be heard in the voices of many otherwise opposed leaders. Outspoken federalist Noah Webster urged that, “…it is necessary to frame a liberal plan of policy and build it on a broad system of education” (Webster, 1785) while the anti-federalist Thomas Jefferson reflected that, “I think by far the most important bill in our whole code, is that for the diffusion of knowledge among the people. No other sure foundation can be devised, for the preservation of freedom and happiness” (correspondence to George Wythe, 1786; see also Jefferson, 1778). The radical liberal Tom Paine advocated in his Rights of Man (1777) for general support for education in order to foster each individual’s lifelong pursuit of knowledge and understanding, preserving thereby a fully participatory, democratic form of self government; thirty years later, conservative activist Benjamin Rush would advocate for public education in his home state of Pennsylvania, suggesting that, “…one general, and uniform system of education, will render the mass of people more homogeneous, and thereby fit them more easily for uniform and peaceable government” (1806, p. 5). These five voices represent distinctly separate philosophical and political viewpoints, sharing in common only the struggle to transform Lockean theory into working democratic institutions, a struggle that led to the development of the formative logic behind the establishment of American public education.

Belief in the democratic imperative proved to be a durable driving force in American education. By the 1820’s, policy makers such as James G. Carter, ‘Father of
Normal Schools,’ were speaking out against educational inequality, decrying a wealth-based system of education that had led to one set of schools for the rich, another for the poor (Carter, 1826). Carter lamented that the educational inequality was, “sapping the foundation of the ancient democracy” (quoted in Hinsdale, 1900, p. 30). Beginning in the 1830’s, Horace Mann adopted the democratic imperative for education as a rallying call for public support for and organization of public schools. From Mann’s perspective, the civilizing and equalizing forces of education made it the single most critical service in the establishment, maintenance, and preservation of democratic society (Spring, 1990; see also Cremin, 1980). Mann worked tirelessly to establish a system of governmentally operated schools that could ensure to all children, regardless of income or geography, a free, comparable, or “common,” education (Wells, 1993).

Mann’s common school movement gained momentum through the 19th century. The states empowered local boards of education to raise taxes to support K-8 and later K-12 education (Cremin, 1980). Passage of compulsory attendance legislation further bolstered the movement (ibid.). Constitutional conventions at the state level framed the need for ‘thorough,’ ‘efficient,’ or ‘uniform’ state systems of public education, with funding and administration to be overseen by the legislative branch (e.g., Goldstein, Gee, & Daniel, 1995). In Ohio, for example, the Constitutional Convention of 1851 led to adoption of Section 2 of Article VI of the Ohio Constitution, which reads in part:

“The General Assembly shall make such provision, by taxation or otherwise, as, with the income arising from the school trust fund, will secure a thorough and efficient system of common schools throughout the state […].

22
By 1900 a system of common schools was in place across the country, funded almost exclusively through local effort by means of property tax (Swanson & King, 1997). This basic system of educational service delivery, organized and managed by elected, local boards of education, and funded based on local, voter approved property tax appropriations, had come to represent the heart of the political and economic life of American public education.

2.2 The struggle for educational equity

Nearly seventy years after James Carter had voiced his concerns regarding wealth-based inequalities in educational opportunity, Ellwood Cubberly in 1905 published the first systematic, descriptive study of educational inequality. So chronic was the problem of wealth disparities that poor schools in most states could not even raise sufficient funds to operate for the full school year. In one township in Perry County the school was forced for lack of funds to close its doors after just 60 days (p. 26). Cubberly’s call was not for absolute equalization, but rather equalization up to some point determined by the state to represent an adequate educational opportunity. “Theoretically all the children of the state are equally important and are entitled to have the same advantages; practically this can never be quite true” (p. 17). Operating from this proposition, Cubberly nevertheless saw a specific state obligation to ensure a minimally adequate education irrespective of community wealth. “The duty of the state is to secure for all as high a minimum of good instruction as is possible, but not to reduce all to this minimum […]” (ibid.). Thus, Cubberly’s work offered a conceptual framework that placed equity concerns within the context of educational adequacy.
George Drayton Strayer, a colleague of Cubberly’s at Teacher’s College, published his doctoral thesis in the same year as Cubberly (Strayer, 1905). Whereas Cubberly’s work was purely descriptive in nature, Strayer offered a statistical analysis in which he measured the degree of variability in educational expenditures in relation to the observed patterns of central tendency (pp. 9-10; pp. 116-123). In this regard Strayer was the pioneer of the scientific measurement of educational equality. Strayer acknowledged the limitation of using expenditures as an indicator of educational opportunity, noting that standardized educational achievement scores would offer a more appropriate measure. With naïve optimism, Strayer suggested that such measures might be available within ‘a year or two.’ Little did he realize that nearly 80 years would pass before the focus on expenditures would finally give way to the direct measurement of achievement equity, as provided here in the present study.

Working within the context of his own work on expenditure data, Strayer recommended a solution to the wealth-based disparities he had observed. His ideas for a state-equalized, minimum level of educational funding across school districts (pp. 173-176) was subsequently incorporated into the New York state funding system (Strayer & Haig, 1923). Refining and promoting his ‘foundation funding’ model (e.g., Strayer, Haig, 1923, Mort, 1924), states across the country adopted the approach as a means of equalizing educational opportunity in the spirit of seeking to fulfill the democratic imperative (Swanson & King, 1997).

Efforts to equalize educational funding notwithstanding, educational disparities remained, not merely with respect to economic status, but also in terms of race, sex, ability, and English proficiency. Beginning in the 1950’s, these long-standing issues of
educational equity were taken up, first in a series of federal Civil Rights cases and followed later by a series of state-level, school-funding lawsuits. This litigation, based on the Equal Opportunity clause of the 14th Amendment of the U.S. Constitution, began with the watershed suit of *Brown v. Board of Education of Topeka, Kansas* (1954) in which the Court sided with plaintiffs seeking equal educational access based on race. Equity litigation following *Brown* diverged in many directions. Equity in educational access based on physical and mental ability began with *PARC v. Commonwealth of Pennsylvania* (PARC, 1972) and led to P.L. 94-142 (1975), also known as the *Individuals with Disabilities Education Act* (IDEA). Equal access to education based on sex is traced from Title IX of the *Education Amendments of 1972* to the seminal federal case of *Cannon v. University of Chicago* (1979). A fourth strand in the struggle for equal access to education includes the legislation and litigation affording equal access for students with limited English proficiency (LEP), flowing from the *Bilingual Education Act* (1968), the *Equal Educational Opportunity Act of 1974*, and the delineation of language equity provided in the Supreme Court ruling on *Lau v. Nichols* (1974).

A fifth line of litigation for educational equity, growing out of the Civil Rights movement, focused more generally on the issue of educational inequality as indicated in overall disparities in fiscal resources for education. This line of cases flowed from the California case of *Serrano v. Priest* (Serrano, 1971), in which the California Supreme Court overturned the state’s system of school funding. When the issue of fiscal equity came before the U.S. Supreme Court, in the case of *San Antonio Independent School District v. Rodriguez* (1973), the majority ruled that plaintiffs were not entitled to federal relief, leaving the issue as a matter for the state courts to decide. State outcomes were
disparate, ranging from judicial inaction, as in *Danson v. Casey* in Pennsylvania (Danson, 1979), judicial-legislative negotiation for improvements, as in the long series of suits in Ohio, from *DeRolph I* (1997) through *DeRolph IV* (2002), to judicial mandates for complete overhauls of state funding, as called for, e.g., in California (*Serrano*, 1971) and Kentucky (*Rose v. Kentucky*, 1990).

Throughout the period of Civil Rights litigation, the overriding operational objective was equalization of educational opportunity by means of equalization of educational inputs, whether defined directly in terms of dollars or in terms of the goods and services purchased by the district or school (e.g., facilities, teachers, and books). Although the U.S. Supreme Court acknowledged that “education is the very foundation of good citizenship” (*Brown*, p. 492), the Court would also rule that, in the case of educational equality for students with disabilities, the IDEA “was more to open the door of public education to handicapped children … than to guarantee any particular substantive level of education once inside” (*Hendrick Hudson District Board of Education v. Rowley*). Civil Rights litigation addressed many of the structural issues of educational equity, and the victories were real and tangible in terms of ensuring at least some level of equalized access to education. Importantly, however, the law did not ensure equality of education from locale to locale nor did the law provide guarantees regarding the overall quality of education. The fight for equality in education had opened the schoolhouse door to all students but was not succeeding in leading the states to the fulfillment of the democratic imperative for education.
2.3 Achievement equalization

The first 100 years of the nation’s history witnessed efforts to devise and build a system of primary and secondary education capable of fulfilling the democratic imperative to prepare all of America’s children for participation in democratic society. Most of the following 100 years focused on the struggle to equalize this basic opportunity. Although the language in that struggle has appealed to the results and benefits of education, the basis of the actual disputes since Brown has tended to consist of challenges to unequal access to the public school and the resources provided there.

In the background of the equity litigation, a public policy trend slowly took shape in which educational achievement, specifically standards-based measurements of student achievement became the benchmark for educational quality and educational equality. This trend, commonly referred to today as the standards-based reform movement (e.g., Fuhrman, 2003, p. 7), yielded in 2002 a national policy on educational achievement equity though the requirements of the revised ESEA. This section reviews the literature describing and documenting the evolution of the standards-based reform movement, from its origins in the 1960’s through passage of the 2002 ESEA reauthorization, and concludes that, in the context of the democratic imperative for education, and in the spirit of educational equity litigation since Brown, the new national policy framework for student accountability offers a meaningful starting point for an output-oriented research agenda addressing the measurement of inequality in educational opportunity.
2.3.1 Testing and educational attainment

In broad concepts, the standards-based reform movement, can be traced back to the middle of the nineteenth century. As part of an effort to measure achievement in the Boston City Schools, Horace Mann and Samuel Gridley Howe developed a test that, in Mann’s words, could determine, ‘whether the pupils have been faithfully and competently taught’ (Mann, quoted in Rothman, 2001, p. 420). The New York State Regents Examinations were first administered at the elementary level in 1865 and at the high school level in 1878 (Britell, 1980, p. 25). In 1929 the Iowa Test of Basic Skills was published which, when combined in the 1950’s with E. F. Lindquist’s electronic scoring technology, set the stage for a comprehensive program of student assessment (Rothman, 2001).

In concert with these technological advances, public concerns were rising regarding the condition of public education (Bestor, 1953). In 1965, sounding a call for accountability that echoed Mann’s sentiments of a century before, Senator Robert Kennedy of New York won legislative support for including in President Johnson’s proposed Elementary and Secondary Education Act (ESEA) a requirement that all programs receiving federal funding under the Act annually evaluate the academic effectiveness of the programs (Popham, 2001, p. 5). But while the Democratic senator was retracing a well-worn trail between testing and achievement, the statutory provision he championed in a federal funding package opened a new path connecting achievement and funding. The implications of this new policy would soon become evident: “If standardized achievement tests could ascertain the effectiveness of ESEA-funded, basic
skills-focused instructional programs, they could be employed to evaluate the success of other types of instructional programs as well” (p. 6).

From a federal mandate for educational accountability among a narrow set of programs to a statewide mandate for K-12 education was only a step; in 1969 the Florida legislature, with little policy analysis or deliberation adopted the first statewide system of minimum competency testing (Amrein & Berliner, 2002). The basic underpinning of competency testing is that, while student ability is not uniform, all students can attain basic academic proficiency (Cohen & Haney, 1980; Popham, 2001). Gorth (1979) defined minimum competency programs as including two parts: “1) the presence of an explicit standard for determining acceptable performance; and 2) the use of test results to make decisions about individual students” (p. 11). One of the distinguishing characteristics of minimum competency testing is the operationalization of a pass/fail standard and the implication of academic consequences for failure, ranging from placement in remedial courses to denial of a high school diploma (ibid). Assuming the educator’s role includes providing instruction and fostering adequate student effort, all students should be expected to demonstrate a clearly defined, minimum level of academic achievement (e.g., Cohen & Haney, 1980).

From its inception in the 1960’s, minimum competency testing spread across the country within a decade. One feature of this expansion was its geographic coverage – from Florida to California, from Massachusetts to Oregon, by 1979, 32 states had adopted some form of minimum competency testing program (Pipho, 1979). Within two years, six more states had joined the group (Lazarus, 1981, p. 3).
Revelation at the end of the 1970’s that SAT scores had declined for the decade, from 478 to 429 on the verbal and from 502 to 470 on the math portion of the test, shocked educators and policy makers (Popham, p. 425, quoting W. Wirtz, On further examination, 1977, see also Ravitch, 1993). The College Board, the developer of the SAT, assembled a blue-ribbon commission to examine the trends in the data and to determine whether the numbers indicated real declines or if the changes could be attributed to other variables such as the expansion of the test-taking pool or changes in the demography of that population (Ravitch, 2003, p. 4). The commission confirmed the worst suspicions: initial declines in the 1960’s were associated with demographic changes, but the more rapid declines in SAT scores during the 1970’s the panel attributed largely to school-related effects. Students’ course taking: academic core courses in composition, literature, and history were being displaced by electives such as cooking and film making (ibid). Absenteeism, social promotion, and low expectations of student productivity also contributed (ibid). All of these concerns were embodied in the commission’s assertion that changes in the public K-12 curriculum had led to general lowering of standards (ibid).

Policy makers reacted by calling for a renewed focus on core achievement and for testing to verify that achievement. The growing sentiment for standards-based reform and accountability is illustrated in the widely publicized and discussed proceedings of the 1977 Quality of Education hearings before the U. S. Senate (Lazarus, 1981), during which Admiral Hyman Rickover, presaging the reforms of the 1980’s and 1990’s, recommended development of an integrated set of grade-appropriate content standards and minimum competency tests and that the results of those tests should be publicized by
school, district, and state report cards (ibid., pp. 62-63). By the end of the 1970’s policy
makers nationwide had embraced minimum competency testing. The following decade
would witness a sharp increase in support for standards-based reform and for more
thoroughly coordinated state efforts to align curriculum, instruction, and assessment.

2.3.2 *A Nation At Risk* and the decade of systemic reform: 1983-1993

Two major events occurred in 1983 that would increase the pace of state-
mandated assessment and accountability. First, the Florida Supreme Court ruled in favor
of Florida’s policy of denying high school diplomas to students who had not passed the
state’s competency tests (Rothman, 2001, p. 430). Although the decision was not binding
on the seventeen other states that also tied student assessment to promotion, the decision
sent a signal across the country that high-stakes testing was a legally viable policy option.

The second major event in 1983 was publication of *A Nation at Risk* (1983),
warning of a “rising tide of mediocrity.” *A Nation at Risk* spurred a reformulation of the
minimum competency efforts of the 1970’s, leading educators and policymakers to
advocate high standards rather than minimum standards for all students (Cohen, 1996;
Warren, 1990). Underlying this push for excellence was a growing sense among
reformers that the path to achieving high standards would involve not merely the
articulation of standards and a system of assessments geared to those standards, but also
an alignment of curriculum and instruction. This combination of wholesale educational
alignment was characterized as ‘systemic reform’ (Cohen, 1996, p. 100). By 1992 over
two thirds of states had embraced the principles of systemic reform, setting the stage for a
new federal mandate in the 1994 reauthorization of ESEA requiring all states to develop challenging standards and aligned assessments by the 2000-2001 school year.

2.3.3 The nationalization of standards-based reform: 1994-2002

In 1989, the National Governor’s Association (NGA), at the request of President Bush, developed a set of national goals for education reform (Campbell, 1995, pp. 78-79). Republican Governor of Iowa, Terry Branstad, and Democratic Governor of Arkansas, Bill Clinton, led the task force. For the next five years a diverse set of stakeholders, ranging from national business leaders to the American Federation of Teachers (AFT) worked to forge the consensus that culminated in March 1994 with Congressional approval of Goals 2000: Educate America Act (Ambach, 1995). The two leading goals of the legislation, as described by one of Goal 2000’s leading backers, G. M. Ambach, Executive Director of the Council of Chief State School Officers (CCSSO), were to raise performance of all students and focus student and teacher effort and achievement through the articulation of content standards (what students should know and be able to do), performance standards (minimum achievement levels to be demonstrated by grade level), and opportunity to learn standards (the conditions necessary for delivering the content standards and meeting the performance standards) (ibid, pp. 68-69).

A distinguishing characteristic of Goals 2000 was its commitment to equal opportunity, embracing thereby the democratic imperative for education. Previous standards and testing reforms implied adequate education for all students, but Goals 2000 sought to make explicit that challenging content and expectations for high performance
should be part of every child’s education (O’Day, p. 101). In O’Day’s words, “if certain
knowledge and skills are deemed necessary for productive citizenship, they must be
available to all future citizens” (ibid). The link to achievement is made by President
Clinton who wrote to South Carolina Governor Campbell, “Because I believe so strongly
that every child can learn, I believe that actual student performance is the best measure of
the extent to which equal opportunity […] has in fact been achieved…”(quoted in
Campbell, 1995, p. 89).

Building and expanding on the ideas of Goals 2000, the 1994 reauthorization of
the ESEA, labeled the Improving America’s Schools Act (IASA, 1994), added federal
reinforcement to the nationwide trend linking content standards, state assessments, and
school accountability. Where Goals 2000 provided several million dollars in federal
assistance to help drive systemic reform efforts in the states, ESEA represented over ten
billion dollars in federal subsidies and program grants (Jennings, 1995b, p. xvi). The
statute required states receiving federal aid under ESEA to set up assessment systems
aligned to academic content standards and to demonstrate measurable progress toward
raising the academic achievement of all students.

2.3.4 No Child Left Behind Act of 2001: National mandate for achievement equity

In 1999, the ESEA was again up for reauthorization. The Clinton
Administration’s recommendations (Office of the President, 2001) underscored the need
for advancing the standards-based reform movement in education, pointing to evidence
from NAEP, the National Assessment of Educational Progress (p. 110, quoting NCES,
1998), that math and reading scores had increased among nine year olds in high poverty
schools and similar evidence from the states that the three-year trend in the proportion of students meeting state standards had increased (p. 110, quoting U.S. Department of Education, 1999). Efforts to advance the recommendations failed, however, as Democratic President Bill Clinton and the Republican-controlled Congress were unable to reach agreement on planned changes (Robelen, 1999).

The abortive *Educational Excellence for All Children Act of 1999* called for expansion of the 1994 mandate for testing. The new requirement would have directed states to publish annual state, district, and school report cards, to disaggregate results by race, sex, ability, economic status, and English proficiency, and for each state to ensure that all students met the state-defined proficiency standards within ten years of Congressional approval of the new law (Robelen, 1999). In essence, the Clinton plan, if passed, would have succeeded in federalizing the accountability movement, drawing together the individual threads of the past forty-five years in equity and accountability development, extending the 1994 mandate for standards and testing by incorporating in addition a mandate that students in each of the named subgroups meet the academic achievement standards.

Election in 2000 of Republican President George W. Bush, far from leading to a new direction in federal policy, resulted in renewed efforts to pursue the Clinton plan. Under the new name of the *No Child Left Behind Act of 2001*, President Bush signed the bill into law in January of 2002 (Robelen, 2002). Through three Democratic and five Republican presidential administrations, what had begun with Senator Robert Kennedy’s efforts to leverage educational improvement through the ESEA had grown and developed
into a systematic framework for educational accountability that ensured a standard set of achievement measures within each state.

In regard to the measurement of equity in student achievement, George Strayer’s dream was at last made possible. The standards-based reform movement has woven together the separate threads of the equality struggle. With the NCLB requirements for state accountability systems to ensure that achievement standards apply not merely to the general student population, but separately and equally to specific subpopulations of students as defined by race, sex, ability, economic status, and English proficiency, the standards-based reform movement offers an output orientation to the question of educational equality. A key question to whether these policies can indeed inform the status of an educational system with respect to achievement equality is whether such equality can be measured in a systematic and meaningful way. The next section addresses this question, tracing from current tools and past studies a methodological framework for the measurement of achievement equity.

2.4 The measurement of educational equality

This section begins with a review of the prevailing methods of measuring equality in educational achievement, namely school and district report cards. The discussion highlights the limitations of test pass rates as indicators of changes in the distribution of achievement. For a more systematic approach to the quantification of the comparative equality of distributions the section turns to the income equality literature. The review demonstrates that the methods developed in the income equality literature have been successfully adapted in the education finance literature for use in the measurement of
fiscal equality in education. This adaptation suggests that, by measuring comparative equality among distributions of achievement scores, the econometric methods from the income inequality literature offer the means for evaluating achievement equality in a more systematically thorough way than report card data. The section presents a preliminary framework for a research agenda focusing on the quantification of achievement equality by which researchers might better inform questions regarding the distribution of student performance, progress within and between schools toward attaining state and federal standards, and whether, or to what degree, sought-for convergence among subpopulations of students is in fact occurring, thereby informing educators and policymakers regarding the achievement aspect of the condition of educational equity and our progress toward the fulfillment of the democratic imperative for education.

2.4.1 Current practice: School report cards as measures of achievement equality

The most common tool for evaluating overall student performance under the new accountability legislation is the ‘school report card’ (Education Week, 2004). Although contents vary by state, the core data consist of student pass rates on the state’s competency assessments. School report cards are sometimes published at different levels of aggregation, such as district (e.g., the Ohio district report cards) or county report cards (e.g., Maryland report cards). The term ‘school report card’ is used here as a general term encompassing these multiple reporting levels.

School report cards were initiated in the 1980’s and 1990’s during the systemic reform movement following A Nation At Risk (Schwartz, 2003). Early adopters included
states as diversely situated as Texas, Maryland, Indiana, Ohio, and Massachusetts (ibid., pp. 146-147). By 2004, in the wake of the 1994 and 2002 ESEA reauthorizations, 43 states were publishing school report cards (Education Week, 2004). Ohio’s school district report card offers an example. Mandated by the Ohio Legislature in 1996 \((Am. Sub. S.B. 55)\), the report card provides fiscal, demographic, student achievement, and discipline data (e.g., Ohio Department of Education, 1997). The achievement data include pass rates on state proficiency exams in 4th, 6th, 9th, 10th, and 12th grades in the academic content areas of reading, writing, mathematics, citizenship, and science.

Academic progress is determined based on these pass rates, with increased pass rates associated with overall improvement in student achievement in a particular content area and grade level.

As a tool for evaluating inequality in student performance, however, report cards present certain difficulties. An examination of achievement equality implies an examination of the distribution of achievement and a determination of whether the distribution of student achievement is increasing while at the same time differences between students are shrinking. With respect to student score distributions, the report cards provide only the proportion of students whose scores meet or exceed the cut score on a given test. The distribution above or below the score might change substantially from year to year, but none of this is captured in report card data. It is possible, for example, that the pass rate for a group of students might increase while student scores in the bottom quintile of that group might actually decrease. Whereas the report card would suggest improvement, an examination of quintile means would reveal a decline in
achievement equality (Rector & Hederman, 1999, offer a detailed critique of such measurement problems using income data).

A systematic examination of achievement equality requires a different set of tools than the report cards offer. The equity measurement literature provides an array of such tools. As described above, the purpose of equality measurement is the quantification of the degree of similarity or ‘sameness’ in a given distribution of some variable. In economics, the variable in question is generally described in terms of personal or national incomes (e.g., Sen, 1997). In education finance, the variables of interest are typically per-pupil educational income, expenditures or some modified version of the two (e.g., Berne & Stiefel, 1984). As Strayer (1905) initially proposed, and as Odden, Berne, and Stiefel (1979) and Berne and Stiefel (1984) reiterated, an output-oriented measurement of the equity of educational opportunity might be devised based on the distribution of student achievement scores. The following sections will review the origins and development of the equality measurement literature, the subsequent adaptation of those measures in the school finance literature, and finally some recent refinements in the economics literature addressing the problem of within- and between-groups decompositions of certain equity measures.

2.4.2 The measurement of income equality

The analysis of income inequality emerged in the 1880’s (e.g., Goschen, 1887; Atkinson, 1888). This early work explored such problems as movements of persons between income brackets (Wolf, 1892) and to changes in numbers of individuals within fixed brackets of wealth (Ely, 1903). Owing to the divergence of results yielded by such
alternative measurement strategies, George Holmes (1892) suggested the implementation of a multiple measurement approach, including the mean, the range, and the distance between median wealth and median number of wealth owners, in order to develop a more comprehensive understanding of the precise nature of changes within any given income distribution over time (p. 141). As the field of study evolved, additional measures such as Sir Arthur Lyon Bowley’s interquartile range statistic (1902, p. 136), adapted as an income inequality measure by Thomas Adams and Helen Sumner (1905) and later by Bowley himself (1920), were incorporated.

Contributions to the conceptual framework surrounding the measurement of equality included Adams and Sumner’s discussion of statistical and conceptual caveats to students (pp. 533-547). Their warning that ‘[…] the common method of presenting statistics upon this subject utterly disguises those facts which the student most desires to know” (p. 534) was directed both at the available economic data and at the methods employed in the evaluation of the distribution of those economic data. Addressing the question of whether the examination of equality bears any relevance to research, the authors are forthright in their claim:

“Back of the question of fact, lies the question of ideal: is the demand for a more equal distribution of wealth justifiable? The disposition of the writer is to answer this question with the strongest affirmative possible. Not only is equality a sound social ideal, but, properly understood, it is the only sound ideal. No one can deny that the existence of widely separated classes, with enormously disproportionate economic and political power, creates problems that threaten the peace and prosperity of the body politic. No one can doubt that the successful operation of a democratic government is facilitated by the homogeneity of its citizenship” (p. 544).
Extending this emphasis on democracy and incorporating the democratic imperative for education and equal educational opportunity, Adams and Sumner assert that equality, “can be secured, permanently, only by slow-paced change making for the more equal distribution of education and opportunity” (p. 545).

Max Otto Lorenz conducted the first general critical review of measures, in which he drew together the early body of work on the measurement of income inequality (1905). Lorenz also offered a statement of the basic problem of income inequality: “There may be a wide difference of opinion as to the significance of a very unequal distribution of wealth, but there can be no doubt as the importance of knowing whether the present distribution is becoming more or less unequal” (p. 209). Replacing the word ‘wealth’ here with ‘educational opportunity’ captures the problem in light of the need for understanding the condition of educational equity.

In addition to Lorenz’s summary of previous research, he offered a new technique for comparing the distribution of income proportions (as contrasted with absolute amounts). Ranking income holders by income percentile, from lowest to highest, and plotting the cumulative percentile of total income held by each successive percentile of income holder, Lorenz could plot any given distribution. Lorenz then compared the observed distribution to a conceptual distribution in which each percentile of income receiver held a perfectly proportionate share of total income such that 10% of income holders would have 10% of the total income, 20% of holders would have 20% of the income, and so forth up to 100%. Plotting the two curves, the observed curve would always appear as a concave curve lying somewhere below a forty-five degree line representing the conceptual distribution of equal incomes.
Hugh, Baron Dalton substantially expanded Lorenz’s work, offering both an expanded array of techniques but also, and perhaps more importantly, a set of principles by which to comprehend shifts in an income distribution, and thereby providing a more compelling theoretical basis for understanding relationships between incomes (1920). Following a suggestion by Pigou (1912, p. 24), Dalton proposed the condition that a transfer of income from a richer to a poorer person, so long as that transfer does not reverse the ranking of the two, will result in greater equity (Dalton, p. 351, citing Pigou). This principle has come to be known as the Pigou-Dalton principle (Sen, 1973; Litchfield, 1999).

Dalton offered a theoretical proposition of a positive functional relationship between income and economic welfare and that economic welfare increases at an exponentially decreasing rate with increased income, expressing the relationship in the form: \( dw = dx / x \) where \( w \) and \( x \) represent, respectively, economic welfare and income. Accepting this proposition, Dalton proceeds to the conclusion that maximum social welfare under such a function is achievable only when all incomes are equal.

Dalton then examined the measurement of inequality with respect to this functional form, offering two measures of his own in addition to measures previously suggested or developed by Pareto (1897), Bowley (1902), Money (1906), Lorenz (1905), Gini (1912), and Pigou (1912). Dalton’s investigation included the first English use of Gini’s measure of mean difference (Gini, 1912), which would later come to be known as the Gini coefficient (e.g., Dalton, 1920).

Dalton’s synthesis produced two important conclusions. First, the utility of a measure is dependent upon the functional form of the economic welfare function of
interest. Second, any given measure and potentially all conceivable measures may offer only limited information with respect to particular variations within a given distribution. With respect to the first of these conclusions, Dalton demonstrates that under the premise of a one-to-one relationship between income and economic welfare, a uniformly proportionate increase in an array of given incomes (e.g., increasing each income by 10%) will result in greater inequality whereas, under the premise of Dalton’s asymptotic functional form, the same proportionate increase would lead to diminished inequality (pp. 355-6).

The second of Dalton’s conclusions is revealed in the observation that, of the measures he examined, none was capable of distinguishing more than one of three potential types of changes across an income distribution, i.e., proportionate additions to incomes, equal additions to incomes, and proportionate additions to persons. At most each of the measures detected two of the three. As an example, a measure such as the absolute standard deviation (the sum of squared deviations about the mean divided by the number of observations) will indicate greater inequality in the case of proportionate additions to incomes but will remain unchanged in the case of equal additions to incomes, while the relative standard deviation (the absolute standard deviation divided by the arithmetic average) yields information on the opposite cases. (Dalton, 1920).

Following the publication of Dalton’s paper, methodological strategies in the field of income inequality remained substantially stable for nearly four decades. In the 1960’s, Henri Theil adapted the measurement of entropy to the study of probability in information theory (Theil, 1967). As unrelated as this connection might appear to be to the study of inequality, Theil had introduced a powerful new analytical tool. Considering
a perfectly unequal distribution of incomes, i.e., all income in the hands of a single entity, as a condition of perfect order (zero entropy), and a perfectly equal distribution of incomes as a condition of perfect disorder (maximum entropy), Theil recognized that the entropy measure offered a direct measure of inequality for any distribution (pp. 94-96). Theil simply subtracted the observed entropy of an income distribution from the maximum entropy possible for that distribution to arrive at a measure of income inequality (ibid.).

Theil’s measure satisfies the Dalton-Pigou requirement that a transfer from a richer to a poorer person leads to a reduction in inequality, or in this case increase in entropy (Sen, 1997). However, as Sen cautions, although the Theil measure is intuitively attractive, it is essentially atheoretical in nature (pp. 35-36). As with other descriptive measures, e.g., range, variance, coefficient of variation, this is an interpretive rather than technical consideration, requiring the analyst to translate the results with respect to a given welfare function rather than simply accepting a ‘less is better’ assumption; gains made at the lower end of a distribution (i.e., among the poor) might be of greater theoretical importance, in which case even small changes resulting from that end of the distribution might be considered more relevant than larger changes at the upper end of the income distribution (i.e., among the rich).

A particularly attractive aspect of Theil’s measure is that, unlike the Gini coefficient, it can be decomposed into within-groups and between-groups components, analogous to the decomposition of effects in the analysis of variance (Sen, 1997; Conceição & Ferreira, 2000). A peculiar feature of Theil’s measure is the fact that it implies a social welfare function of the form $x_i \log(1/x_i)$, suggesting steeply declining
marginal utilities as income (or education) increases (Sen, p. 36). From an educational perspective this suggests that the difference between literacy and illiteracy is of greater marginal utility than between a master’s degree and a doctorate. Thiel’s measure must be interpreted in light of this underlying functional form (Sen, 1997).

Anthony Atkinson who in 1970 published a revised version of Dalton’s inequality index, offered the most recent measurement technique that has appeared in the educational equity literature. Wedgwood (1939) had presented an empirical application that Champernowne (1953) used to refine the technique, defining the concept of ‘equally distributed equivalent income.’ Atkinson (1970) extended Dalton (1920), Wedgwood (1939), and Champernowne’s (1953) approaches, modifying Dalton’s basic technique to be independent of effects due to linear transformations of the data. Atkinson, recalling Dalton’s (1920) essay, was convinced that the interpretation of any measure of income inequality implies some concept of social welfare and he carefully specified the social welfare function embedded in his measure. This is a caution to which we will return in chapter six as we discuss the interpretive implications of the various measures.

2.4.3 The measurement of educational equity: From Cubberly to Berne & Steifel

As discussed under section 2.2, above, Ellwood Cubberly offered the first general description review of educational equity (1905). Cubberly’s contribution was to begin to document differences not simply on a selective basis, as Jefferson and Mann, for example, had done, but on a statewide basis, comparing all districts in the states of New York and Massachusetts. Cubberly examined district proportionate taxing power and
expenditure levels. His measures included ranks and percentages both for district totals and per-pupil ratios (ibid).

George Strayer, also introduced in section 2.2, above, was the first to offer a systematic approach to the measurement of educational equality, incorporating for the first time certain measures of association, i.e., the Pearson correlation coefficient, into the generalized quantification of educational equity (1905). Strayer’s analysis also made wide use of the median rather than the mean in light of the median’s insensitivity to skewness, a weakness of the mean (ibid). Strayer provided a substantially more comprehensive and rigorous methodology than Cubberly but, interestingly, both men turned their attention to other problems in education finance and never developed these initial contributions any further.

Following Cubberly and Strayer, the search for methodological refinements in the study of educational equity led analysts to borrow various methods developed by economists for use in the analysis of the equality of income distributions. In the wake of the Serrano (1971) and Rodriguez (1973) litigation of educational equality, researchers turned their attention to the measurement of equality in education, particularly in terms of the distribution of fiscal resources across districts (Mushkin & McLoone, 1965; McLoone, 1974; Grubb & Michelson, 1974; Odden, Berne, & Stiefel, 1979; Berne & Stiefel, 1984; Odden & Picus, 2000). Borrowing the tools developed in the income equality measurement literature, educational economists conducted analyses ranging from national studies (e.g., President’s Commission on School Finance, 1970; Brown, et al., 1977) to single state studies exploring fiscal equality within individual states, beginning with Guthrie, et al. (1971).
By the end of the 1970’s, the fiscal equity literature had matured to the point where researchers were able to describe and define a systematic conceptual framework (Odden, Berne, & Stiefel, 1979). Further refined by Berne and Stiefel (1984), the educational equity framework is designed around three foundational concepts, including variable classification, objective function of interest, and level of analysis. The equality of education might be defined or measured in terms of any among numerous variables of possible interest. Students of educational equity have adopted three general classifications of educational variables following an input-output model of educational production (Odden, Berne, & Stiefel, 1979; Berne & Stiefel, 1984). Inputs include revenues, expenditures, and purchased resources. Revenues are classified by source. The basic array of revenue sources is constituted of local taxes and fees, and state and federal intergovernmental transfer payments. The local tax base is made up largely of property value and, in a small number of states and districts, local income. Local taxes account for approximately half of all local revenue (Swanson & King, 1997). Given the wide variability in local property wealth and income, these local revenues tend to be highly unequal. State revenues are drawn from the state general revenue fund, which in turn derives mainly from a mix of personal and corporate income taxes plus sales and excise taxes. State transfer payments to school districts are generally distributed in two parts: an equalized, foundation aid portion directed to the general education program and an unequalized, categorical aid portion directed to support of specific services such as special, gifted, and vocational education. Federal revenues are based in the federal income tax and are generally distributed in the form of unequalized, categorical aid.
However, owing to the fact that most federal categorical programs tend to be associated with poverty, the net effect of federal aid allocations tends to be equalizing (ibid.).

As an indicator of educational service or delivery, revenues offer little more than a crude indicator of raw resources. Expenditures, by contrast, represent the actual goods and services purchased by each district and delivered, presumably, to the students. Expenditures can be subdivided into administrative categories such as direct instruction, pupil support, staff support, operations, and capital. This enables refinement of the revenue figures in order to compare, for example, equitability of direct educational allocations minus variable expenditures such as capital and debt service. While such accounting provides substantial additional information regarding the educational systems across districts, there remains the fact that expenditure data are influenced by local and regional price differences. These differences are attributable in part to demand factors and in part to supply factors. In order to establish truly comparable figures across districts, the expenditure figures should be adjusted to eliminate variation associated with supply factors. Unfortunately, it is not possible to distinguish between effects associated with supply factors and those associated with demand factors. As a result, expenditure data, while improving on revenue figures, are not comparable across districts (Berne & Stiefel, 1984).

Purchased resources, the third category of inputs, include such things as the teachers, support staff, administrators, and textbooks acquired by the district to serve the student. Ratios of these goods to pupils, e.g., pupil-teacher ratio, provide measures of resources that are independent of the type of supply factors that confound expenditure data. Data for purchased resources suffer, however, from the fact that they are non-
additive across units, i.e., the pupil-teacher ratio cannot be added to the pupil-administrator ratio in a meaningful way. As a result, purchased services provide a more refined level of information, but provide it in discrete units that make aggregate comparison impossible (ibid.).

In contrast to inputs, outputs refer to the immediate results of education. These include for example student grades, standardized test scores, attendance and graduation rates, and discipline rates. Before the adoption of statewide data reporting systems and accountability programs, researchers lacked common measures of educational outputs. Adoption of state achievement measures and federal accountability policy offered a democratically derived set of student output measures and a conceptual framework in which to place them. This topic is highlighted later in the final section of this review of literature.

The level of analysis in educational equity studies addresses the question of, Equality between what? The researcher might focus on distributions of an educational variable across countries (e.g., Checci, 2000), states (e.g., Guthrie, et al, 1971), districts (Berne & Stiefel, 1984), or schools (e.g., Berne & Stiefel, 1994). If the level of analysis is the school district, for example, then the research will address the question of how equally the educational variable of interest is distributed across school districts.

Fiscal equity distinguishes between three equity objectives: horizontal equity, vertical equity, and equal opportunity. Since achievement equity deals with only the first of these, we describe it here in detail; for a general discussion of vertical equity and equal opportunity, the reader is referred to Swanson & King (1997). Horizontal equity is characterized by the objective of achieving equal shares of the variable of interest across
equally situated members at the selected level of analysis. If the variable of interest is the input variable of expenditures, and the level of analysis is the pupil-weighted school district, then horizontal equity is measured based on the degree to which per-pupil expenditures are equally distributed across equally situated districts. The ‘equally situated’ qualification in this case invokes the idea that if costs differ between districts then the observed expenditures should be adjusted accordingly in order to remove cost effects from the expenditure measure. For input, output, or outcome variables, respectively, horizontal equity would require equal expenditures per pupil (instruction, administration, etc.), equal educational services (number and quality of instructors, number and quality of guidance and other support staff, etc.), or equal support services (equal content and quality of food program, extracurricular program, etc.). From the perspective of achievement equity, we will be interested in horizontal equity because the objective of achievement standards is to raise all students to the standard regardless of student or external characteristics.

As Dalton (1925) and Sen (1973) suggest, and as Odden, Berne, and Stiefel (1979), and Berne and Stiefel (1984) further demonstrate, horizontal equity can be measured by diverse methods, each with its own particular characteristics. Simple measures of horizontal equality include the range, the restricted range, i.e., the absolute difference between the observations at the fifth and ninety-fifth percentiles in a rank-ordered distribution, and the federal range ratio, i.e., the fifth percentile divided by the restricted range (Berne & Stiefel, 1984). While Sen (1973, 1997) dismisses these simple measures of horizontal equity as ‘non-starters’ for obvious reasons, e.g., the range-based statistics offer only the difference between two extreme points (though somewhat less
extreme in the case of the federal range ratio) in a distribution; either of the two points might change without corresponding to a general change in conditions across a distribution. In spite of this shortcoming educational researchers continue to incorporate the measures into reviews of educational equality (e.g., Berne & Stiefel, 1984; Odden & Picus, 2000).

A second group of measures of horizontal equity are based on the variance statistic. Included here are the standard deviation (the square root of the variance), the coefficient of variation, and the standard deviation of logarithms (Dalton, 1920; Sen, 1973; Odden, Berne, & Stiefel, 1979). As with the point measures of horizontal equity, the variance-based measures offer certain limitations. In the case of the standard deviation, reliance on the mean as a fixed point of reference leads to relative differences in equity based on the size of the mean (Sen, 1973). The coefficient of variation, while independent of the mean, treats changes at either end of the distribution equally, whereas most concepts of social welfare would value increases at the lower end of the distributions more highly than decreases at the upper end of the distribution (ibid). The standard deviation of logarithms eliminates the standard deviation’s dependence on the mean and the coefficient of variation’s equal treatment of changes at the high and low ends of the distribution (ibid). The standard deviation is independent of the unit of measure and, by virtue of the logarithmic transformation of the data, yields increased sensitivity to differences at the low end of the distribution. Given these positive characteristics, a concern raised regarding the standard deviation of logarithms is whether we should expect social utility to follow the particular functional form implied by the statistic, an expectation for which no rational basis has been offered (Sen, 1997).
A third group of measures of horizontal equity include three measures discussed in the section above on income equality. The Gini coefficient (Gini, 1912), the Atkinson index (Atkinson, 1970), and the Theil statistic (Thiel, 1967) make up the three distributional measures. We refer to this group of measures as distributional measures of horizontal equity because they incorporate information about each observation in the distribution with respect to all other observations in that distribution. The Gini coefficient, for example, involves a computation of all pairwise differences between scores. A somewhat crude but accessible descriptive approach to understanding these distributional changes is the quintile analysis (e.g., Lynch, 2003) by which the researcher may quickly examine the basic characteristics of a distribution or set of distributions before moving to the distributional measures.

The distributional measures offer several technical and interpretive advantages over the point- and variance-based measures. The Gini coefficient represents the cumulative distribution of scores in relation to the cumulative distribution of subjects (e.g., schools). The Gini has the intuitive appeal, as a ratio varying on the range \((0,1)\), of a meaningful minimum, i.e., all value is concentrated in a single observation, and a meaningful maximum, i.e., all value is equally distributed across all subjects (Sen, 1997). The Atkinson index, as discussed above, offers a distribution wide measure of horizontal equity based on a formally stated social utility function over which the researcher retains control by means of an ‘inequality aversion’ coefficient, \(e\). An appealing aspect of the Atkinson index is this choice of social welfare sensitivity exercised by the researcher (ibid). This choice of \(e\) may, however, become arbitrary in the absence of a strong theoretical rationale (ibid). Whereas the Atkinson index conforms to a formal, though
possibly arbitrary social welfare function, the Theil statistic invokes instead a well understood law of physics, i.e., entropy (Theil, 1967). The application of physical laws in social science is well established, e.g., Newton’s law of gravitational attraction has been widely applied in geography, transportation planning, and city and regional planning. Although social phenomena have been observed to conform, to greater or lesser extents, to these physical analogs, there remains nevertheless the problem that these applications are essentially atheoretical; i.e., there has been no explanation regarding why we might expect the distribution of educational attainment to conform to the entropy function (Sen, 1997).

2.4.4 Within- and between-groups measures of equality

A useful aspect of the distributional measures of horizontal equity, i.e., the Gini coefficient, Atkinson index and Theil statistic, is that these measures can be broken down in such a way that researchers and policy analysts can determine the unique contributions to equity of differences within and between subgroups (Litchfield, 1999). The concept of within- and between-groups inequality is analogous to that of the analysis of variance, in which total variance is the sum of the variance within groups and the variance between groups (Akita, 2003). In his original monograph, Theil (1967) noted the potential for decomposability, but not until Bourguignon (1979), Cowell (1980), and Shorrocks (1982a, 1982b, 1984) was the concept more fully explored and applications developed.

The within- and between-groups approach has been employed in the examination of income equality within and between countries (e.g., Sala-i-Martin, 2002), regions (Conceição & Ferreira, 2000) and gender (Mussard, Seyte, & Terraza, 2003). While
recent efforts have offered revised techniques and applications in the decomposition of the distributional measures of equality (Lasso de la Vega & Urrutia, 2003; Conçeição & Ferreira, 2000; Mussard, Seyte, & Terraza, 2003), these techniques have not yet been adopted for the evaluation of educational equality. The hierarchical nature of educational organization, student subgroups within classrooms, within schools, within districts, within states, suggests the applicability of the equality decomposition approach to educational phenomena, a point that we address in the next section.

2.4.5 The measurement of achievement equity

In retrospect, the democratic imperative for education, the equality movement, and the standards-based reform movement all point to the need for the measurement of educational equality in terms of student achievement. The equality measurement techniques provide the means by which to pursue this problem. A comprehensive system of standards and standards based assessment represents the articulation of an egalitarian standard in education, reconciling, “proved individual differences with the political demands for equality of achievement in a society that espouses equality of opportunity” (Britell, 1980, p. 26). The promulgation of the federal mandate established a national framework for achievement standards for school and student accountability, disaggregated by race, sex, ability, economic status, and English proficiency. This framework represents a new, national policy for achievement equity that lays before educational researchers and policy analysts a new set of problems in the measurement of inequality in educational opportunity, problems based on inequality in educational achievement rather than educational spending.
Interestingly, in spite of this apparent gap in the equality literature, emphasis continues to be placed on educational resources rather than student achievement, e.g., Hirth, (1994), Prince (1997), Murray, Evans, & Schwab (1998), Odden and Picus (2000). We propose here a shift in focus away from educational equity as defined in terms of fiscal resources. The standards-based reform movement, drawing together as it does the equity interests of the educational equity litigation of the second half of the twentieth century, offers educational researchers the opportunity to explore educational equity in terms of the educational achievement.

2.5 Developing an achievement equity research agenda

The steady increase in emphasis on educational outputs, as represented by student-based measures of student achievement, in educational equity policy has created an opportunity to extend the inequality measurement literature in education. While the concepts and methods of inequality measurement clearly anticipate examination of output inequality, and student achievement inequality in particular (Berne & Stiefel, 1984), researchers have not yet turned their attention to this problem. Until the 2002 reauthorization of the ESEA attention to this issue may have been muted by the absence of a national framework or agenda for standards-based accountability. After all, the question of achievement equity can only be addressed with regard to some standard of achievement. The federal mandate for academic equity as embodied in the 2002 ESEA created, at the state level at least, a framework for achievement standards by which achievement equity can be defined. Indeed, the measurement of achievement equity can address certain limitations of test score reporting. That is, the description of student
achievement has been limited to passing rates on the tests with little or no regard for the composition or changes in the distributions of scores. As the preceding discussion has made clear, public policies surrounding achievement accountability and achievement equity are aimed at fundamentally transforming the distribution of student achievement. Though each state’s standards and measures might be unique, the nationwide endorsement of test-based accountability calls for a research agenda examining the degree to which the current methods of summarizing student distributions, i.e., state and local report cards, adequately reveal sought-for improvements.

A second problem with the application of equality concepts to educational achievement is the conceptual difference between the theoretical goal of absolute equality as might be computed for educational resources versus the more limited goal of reduced inequality as embodied in the proposition of minimum achievement standards. Pozner (1992) characterizes the condition in which “some have more and some have less, but there are no have-nots,” as “nonegalitarian equality” (p. xx). From an educational perspective we might turn this around somewhat in describing the condition of a system in which all possess adequate education for democratic participation but in which any particular individual might nevertheless demonstrate a level of achievement different from any other as ‘egalitarian inequality.’ That is, the system Cubberly described when he stated that, “[t]he duty of the state is to secure for all as high a minimum of good instruction as is possible, but not to reduce all to this minimum” (Cubberly, 1905, p. 17) is egalitarian at least to the extent that every individual enjoys equal access to democratic participation. The system is unequal, however, to the extent that individual achievement above the minimum threshold might be expected to vary freely, in relation perhaps to
individual motivation, ability, opportunity, or other personal, group, or environmental factors.

2.6.1 Proposal for empirical study

The literature reveals a gap in our knowledge regarding educational inequality. A shift has occurred in education in which the emphasis on educational funding has been superseded by emphasis on educational achievement. The inequality measurement literature, however, has not yet responded to this shift. Whereas funding was the basis of evaluating progress toward fulfilling the democratic imperative for general education, student achievement has migrated to the center of public debate over educational adequacy. From basic education to equalized opportunities by race, sex, ability, economic status, and English proficiency, the centerpiece of public policy in education at the turn of the twenty first century is student achievement.

Understanding the attainment of achievement based equity goals implies the measurement of the distribution of student performance. The tools and conceptual frameworks of inequality measurement provide the means by which this understanding might be gained. Whereas report cards provide certain pieces of useful information regarding the distribution of student achievement, as measures of central tendency report cards are not adequate tools for systematically understanding distributional inequality any more than average school expenditures. A systematic understanding of achievement inequality among districts, schools, and student groups calls for a research agenda aimed at exploring the quantification of distributional inequality in student achievement. The
The natural starting point is the application or adaptation of known inequality measurement techniques and, as the need arises, the development of new techniques.

The scope of this overarching research agenda will be broad and will include numerous complex components. Studies need to be made of changes in inequality over time, inequality among regions, district types, districts, school types, schools, and student groups. The research must include decomposition of measures in order to enable separate examination of between-groups and within-groups contributions to inequality. Norms must be developed, if possible, to establish statistical and substantive significance. New methods might be developed, if possible, in order to incorporate the criterion-referenced framework of state proficiency cut scores and school and district performance rating criteria.

This study takes the first step toward a new research agenda in the measurement of inequality in student achievement. The methodology incorporates a basic set of inequality measurement techniques defined from the literature and framed in terms of student achievement. A single, large Midwestern school district provides data for an empirical analysis. Examining the distribution of student performance over five years, across fourteen elementary schools, this empirical analysis provides both an illustrative example of the mechanics and interpretation of the measurement of achievement equity.

2.6 Chapter summary

The maintenance of a democratic society relies on an adequately educated citizenry. This proposition implies the principle of educational equity. Educational equity exists when educational achievement is fairly distributed across the population;
from a democratic perspective, educational equity suggests equality of achievement to some level or standard accepted as adequate by citizens acting through democratically representative processes. Some of the earliest concerns with educational inequality were raised by Thomas Jefferson and have been reiterated ever since. Interest in educational inequality has focused mainly on access (civil rights) and support (school finance). Civil rights litigation through the 1980’s focused on equal access to education. School funding litigation has focused on equalizing fiscal resources for school children. Research relating to school funding equity led to the importation from the field of economics a set of techniques for the measurement of inequality. In the 1980’s, as standards-based reform was embraced as a national policy priority, the focus in educational equity shifted away from access and resources to center instead on educational achievement. In 2002, state by state definitions of adequate educational achievement were drawn together under a federal umbrella of educational accountability for student achievement, marking the culmination of the first, federally coordinated, state operated system of educational achievement standards. Whereas inequality measurement literature in education historically focused on educational inputs, the reorientation to outputs calls for the development of an output-based research agenda. Combining the information yielded from standards-based assessments with current inequality measurement methods and concepts, researchers and policymakers will have at their disposal a powerful new approach to measure and analyze the condition of educational equity and progress toward fulfilling education’s democratic imperative. This study represents a first step toward developing the research agenda in the measurement of achievement inequality.
The accountability elements of the current ESEA are numerous and complex, requiring achievement of state standards in multiple achievement domains at the school level by each of several demographic classifications of students. This study examines a cross-section of these elements by examining achievement inequality over a five-year period across twelve elementary schools in one district. The socio-spatial elements are limited in order to accommodate at the same time an examination of multiple inequality measurement techniques. Defining the scope in this way permits a broad, foundational investigation that will readily lend itself to expansion in any one of many directions in future research. Chapter three provides a detailed description of this study’s research questions, hypotheses, and analytical methods.
CHAPTER 3

METHODOLOGY

3.1 Introduction

The purpose of this chapter is to provide a methodological design by which to
address the research questions posed in chapter one, thereby extending the empirical
analysis of educational equity into the sphere of achievement equity. Horizontal equity,
as discussed in the review of literature, above, refers to the degree of similarity among an
observed distribution of values; the greater the similarity of values, the greater the
horizontal equity. In accordance with the research questions, the methodology deals with
the horizontal equity of student achievement. Five contributions are made. First, the
empirical analysis of horizontal equity, traditionally considered from the perspective of
educational expenditures (e.g., Berne & Stiefel, 1984; Odden & Picus, 2000) is
considered here with respect to student achievement as measured by reading proficiency
scores. Second, the horizontal equity analysis, typically applied to the inter-district level
(e.g., Berne & Stiefel, 1984) is applied here at the intra-district level. Third, the analysis
applies recent developments in the decomposition of equity statistics into within- and
between-groups contributions to overall horizontal equity, i.e., Conceição & Ferreira’s
factorial decomposition of the Atkinson index, to the examination of within- and

60
between-schools equity. Fourth, the various equity statistics employed in the foregoing analyses are evaluated with respect to their general functional utility to the analysis of the horizontal equity of student achievement. Fifth, the analysis conducted at the within- and between-schools levels is evaluated with respect to its potential applicability to other levels of analysis, e.g., within and between districts, and to nested designs such as within and between student demographic groups, nested within schools, nested within districts. We will review these five contributions in chapter six, below. We turn now to the methodological design of the study.

3.2 Study period

Data are collected and analyzed for each of six years, 1998 through 2003. Change in performance and equity are examined both where 1998 is treated as a base year against which future changes are measured and, alternatively, where each year is treated independently of the others in the determination of trend across the study period.

3.3 Geographical coverage

The study focuses on a single, large, Midwestern school district. The district operated twelve elementary, three middle, and two high schools. For each of the six study years, average district enrollment was approximately 14,000 students. Racially, approximately 20% of the student body was composed of minority (i.e., nonwhite) students, approximately equal to the state average. Average per-pupil property valuation was approximately $139,000, which ranked at about the 80\textsuperscript{th} percentile in the state. Per-pupil expenditures of $8,000 ranked at about the 71\textsuperscript{st} percentile. Expenditures tend to be
allocated somewhat disproportionately to teacher salaries which, averaging nearly $56,000, ranked in the 97th percentile statewide. The higher teacher salaries are also associated with above average teacher experience, with teachers in the district averaging 15 years compared to a statewide average of 13.

3.4 Student grouping variable

The study design employs one administrative categorical variable, school of enrollment (coded 1-12). The discussion following the analysis addresses the potential for including additional grouping variables such as student’s demographic characteristics.

3.5 Student achievement measure

The study examines the distribution of student achievement based on the scale scores for the 6th grade Ohio proficiency test of reading (OPT). The following section provides detailed information on the background of the test and considerations of the reliability and validity of score interpretations.

3.5.1 Ohio proficiency tests

In 1989 the Ohio legislature mandated the development of statewide proficiency testing as part of a sweeping education reform bill aimed at building standards-based school accountability (H.B. 182, 118th GA). Ohio proficiency test batteries were developed for grades four, six, nine, and twelve. The batteries at each grade are composed of five content specific tests. These include reading, writing, mathematics,
citizenship, and science. The tests are criterion-referenced assessments of student grade-
level proficiency (ODE, 2002).

In accordance with the federal performance accountability mandate, the study will
incorporate results for the test of reading. This selection does not provide comparability
of measurement across states, since Ohio’s proficiency tests are unique to that state.
However, the selection of reading does represent a one of the specific cognitive domains
in which student achievement has been mandated nationwide, and thus offers a common
conceptual basis for understanding one of the student output objectives at the state, local,
and school levels while providing an empirical illustration of how an analysis including
additional measures might proceed.

3.5.2 Reliability

A measurement is reliable if it consistently measures whatever it measures. One
statistical method of evaluating reliability, Cronbach’s alpha, provides an indication of
the degree of internal consistency among items on a test (Hopkins, 1998). A value of
Cronbach’s alpha equal to or greater than .80 suggests a satisfactory degree of reliability.
The Ohio Department of Education reports the reliability of the Ohio proficiency test of
reading for the study years, 1998 through 2003, as ranging between .84 and .87. This
information suggests that the test items used on the Ohio proficiency test of reading
function in an internally consistent manner.

A second indicator of reliability is the standard error of measure, or SEM
(Hopkins, 1998). Since a test only samples a portion of what a student knows regarding a
given cognitive domain, a student’s observed score represents an estimate of that
student’s true score in the sense that the ‘true score’ is the level at which the student would achieve if we could directly measure everything the student knows about the cognitive domain of interest. The SEM indicates the probable degree of variability in observed scores around students’ true scores. The smaller the SEM, relative to the scale of the metric, the more consistently the observed scores will represent true scores. For a given student score, we may be 95% confident that the student’s true score falls somewhere between that observed score and plus and minus $t_{.05}$ times the SEM, where $t_{.05}$ represents Student’s $t$ value taken from a standard table at alpha = .05 and $n = \infty$ (the latter because the number of students taking the test each year is in excess of 100,000 and the value of $t_{.05}$ converges on its asymptotic limit of 1.96 above $n=100$).

The Ohio Department of Education reports the SEM for the reading test for each year of the study period, 1998-2003, as ranging between 11.18, in 1999, and 12.09, in 2002 (Ohio Department of Education, 2004). The results for 1999 suggest that, for a student with an observed score of 200, we can be 95% confident that this student’s true score fell somewhere between 200 plus and minus 1.96 times 11.18, i.e., between 187.64 and 222.36. Since the cut score for the test is 220, this student’s score of 200 and the SEM of 11.18 suggest that the student might either have been proficient or might have been more than 30 scale score points below the cut for reading proficiency. This phenomenon is somewhat exacerbated with increasing distance from the cut score. The Ohio Department of Education (2004, reproduced in Appendix A) reports that a raw score of 1 (i.e., student correctly answers just one item) is associated with a scaled score of 65 while a raw score of 2 is associated with a scale score of 89, a difference of 24 points. By comparison, the scaled-score difference between raw scores of 7 and 8 is 5
points (136 vs. 141). The greater span between scale scores at the extremes illustrates the fact that scores near the extremes are associated with less stability. From the perspective of reliability, these results suggest that the reading proficiency test scores, and pass rates in particular, must be interpreted with caution.

3.5.3 Validity

Validity refers to the inferences that can be drawn from test scores (Messick, 1989). Gullicksen (1950) drew the conclusion that a researcher who does not know what the criterion scores on a measurement mean, cannot know much about the measurement. In a pure sense, a measurement is valid if it measures what it purports to measure and nothing else (Hopkins, 1998). A measurement cannot be valid if it is not reliable because a valid measurement will consistently measure only that which it is designed to measure. A measure may be reliable and yet lack validity if the measurement is consistent but measures something other than what it is intended to measure. Validity is a complex concept encompassing several interrelated facets. Two key facets that relate to proficiency testing are content validity and construct validity. Content validity addresses the question of whether the items are representative of the universe of content defined in the instrument specifications. Construct validity addresses the question of whether the scores correspond to the behavioral domain associated with proficiency in the academic domain.
3.5.4 Reliability of the Ohio 6th grade test of reading proficiency

The reliability of the 6th grade test of reading proficiency has been documented by the Ohio Department of Education (2004) for each of the years used in the study. Table 3.1 presents the results published by the agency.

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean</th>
<th>Reliability</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>222.6</td>
<td>0.86</td>
<td>11.54</td>
</tr>
<tr>
<td>1999</td>
<td>222.2</td>
<td>0.86</td>
<td>11.18</td>
</tr>
<tr>
<td>2000</td>
<td>221.6</td>
<td>0.84</td>
<td>11.73</td>
</tr>
<tr>
<td>2001</td>
<td>226.3</td>
<td>0.86</td>
<td>11.45</td>
</tr>
<tr>
<td>2002</td>
<td>225.5</td>
<td>0.87</td>
<td>12.09</td>
</tr>
<tr>
<td>2003</td>
<td>227.6</td>
<td>0.87</td>
<td>11.68</td>
</tr>
</tbody>
</table>

Table 3.1: Statewide average, reliability, and standard error of measure for the 6th grade reading test

The results suggest that reliability of the 6th grade test of reading proficiency is acceptable, with reliability (Cronbach coefficient alpha) exceeding .80 for each year in the study.

3.5.5 Validity of the Ohio test of 6th grade reading proficiency

The validity of a criterion-referenced test, one designed to measure a clearly defined universe of items, may appear to rest on confirmation of the representativeness of the content of the included items (e.g., Campbell & Fiske, 1959; Osburn, 1968; Ebel, 1974; Yalow & Popham, 1983). This evidence might further be supplemented by a detailed account of the test development process, documenting the domain specifications.
together with expert judgments of both the domain and the item coverage of that domain.

Ebel articulated the requirements of content validity as follows:

The evidence for intrinsic rational validity will consist of an explicit rationale for the test: a written document that (a) defines the ability to be measured, (b) describes the tasks to be included in the test, and (c) explains the reasons for using such tasks to measure such an ability. The explicit rationale indicates what the test is measuring. If that is what the user intends to measure, the test is a valid test for the user’s purposes (1983, p. 8).

The Ohio Department of Education (ODE) employed just such a systematic test development process. The process began with the definition of the purpose and methods for developing and administering the tests. This process included input from nationally recognized assessment experts together with Ohio educators and instructional experts. Content Advisory Committees staffed by Ohio educators then worked with the assessment contractors to develop test specifications, including coverage, length and item formats. Test specifications were based on the Ohio academic content standards. The assessment contractor developed preliminary items. The ODE, the Fairness and Sensitivity Committee, a panel of Ohio educators and stakeholders organized for that specific purpose, and the Content Advisory Committee each reviewed the items for fairness, grade-level appropriateness, and content representativeness. The contractor field tested approved items and provided item analysis results to the two committees for a second round of review. Items and/or test specifications were refined based on the results of the field tests and final forms were produced and, after receiving final approval from both review committees, administered statewide (ODE, 2002). Following Ebel et al, the
ODE’s systematic test development process suggests that the tests are content valid and, by virtue of the empirical analysis of item responses, reliable.

Content validity if taken alone may present logical weaknesses to the extent that validity is understood to be a characteristic of measurements rather than the instruments from which the measures are taken (Messick, 1989; Messick, 1981; Messick, 1980; Messick, 1975; AERA, 1955; Cureton, 1951). Offering an illustration of the interpretation of high scores versus low scores, Messick points out that while high scores may indicate through consistent achievement of representative tasks that a person has acquired the associated domain skills, low scores cannot by themselves shed light on whether, through consistent failure to achieve representative tasks, a person has failed to acquire the associated domain skills (1989, p. 42). The interpretation of low scores is subject to such plausible rival hypotheses of irrelevant test variance as low motivation, anxiety, or sensory handicap.

Plausible rival hypotheses cannot be discounted through an examination of content validity alone. Consequently, validation of score interpretation requires an integrated examination of multiple aspects of validity, incorporating content judgments with empirical analyses and hypothesis testing (ibid; see also Cronbach, 1971). The range of conceivable evidence that might bear on the problem exceeds the feasibility of incorporating it all into a single study, even of exhausting it all in a coordinated research agenda (Messick, 1989, p. 50). The question then is whether, in addition to the ODE’s own evidence of content validation, there exists a body of research that supports, or at least fails to disconfirm, the validity of the OPT.
Various validity studies have been undertaken on the OPT, focusing particularly on the high stakes ninth grade level. The ninth grade level is of particular interest because of its use as a graduation requirement, i.e., a high-stakes test. Since all tests were developed under the direction of a single office following essentially identical procedural systems, it is assumed here that the validity studies of the ninth grade test should be reasonably representative of the validity of the tests at the other grade levels (ODE, 2002).

In a review of the OPT validation research literature, Kim (2002) identified studies focusing on content (Wallick, 1998; Chandler & Brosnan, 1995; Noel, 1994; Gallagher, 1993; Hull & Tache, 1993; Kunicki & Osborne, 1992; Lanese, 1992), response process (Heiney, 1998), and relations to other variables (Gibson, 1997; Green, 1998). To this body of work Kim contributed an analysis of the 9th grade tests’ construct validity as evidenced by internal structure, concluding that four of the tests represented distinct constructs but that the evidence for the writing test was ambiguous, possibly as a result of the test’s particular scoring features (2002).

This study employs results from the 6th grade test of reading proficiency. The 6th grade test is administered only once for each student cohort, is designed to reflect grade-level proficiency, and is intended to be used for determination of student achievement not merely at the cut score but across the full range of scores as is suggested by the use of ‘advanced’ proficiency and ‘basic’ proficiency cut scores. The 6th grade test is the most advanced, cumulative measure of student proficiency available before high school, offering a midpoint measure of elementary preparation and high school preparedness. The choice of the 6th grade test is supported by the fact, although it has not been
subjected to the close scrutiny of the 9th grade test, the development and construction processes were the same as for the 9th grade test and conducted under the auspices of the same technical advisory panel of experts (Ohio Department of Education, 2004).

In conclusion, the body of research provides a broad though limited basis for relying on the validity of interpreting OPT scores as indicative of the degree of student proficiency in the domain of reading. The similarity of design and construction, and the fact that this has been undertaken under the auspices of a single technical advisory committee with a group of collaborating contractors lends support to the likely validity of the tests. Since validation studies have focused on the 9th grade test, and the only construct validation was based on subscale scores, use of the 6th grade results must be taken as suggestive rather than conclusive evidence of student proficiency.

3.6 Population

The study district enrolls approximately 1,000 students per grade level, somewhat more in the early elementary (K-3) and somewhat fewer in high school, due to a slow but steady expansion in residential housing in the district. The study population is made up of students enrolled in grade six at the time of testing in March and who were required to sit for the Ohio proficiency test in reading. Not included in this population were: a) students exempted from testing in accordance with federal standards regarding the assessment of students with disabilities (amounting to approximately 6-8% of students each year); and, b) students waived in accordance with state standards governing the assessment of students with limited English proficiency (amounting to fewer than 1% of students each year).
3.7 Source of data

The source for all data used in this analysis was the district’s office of Research, Planning and Assessment. The office provided students’ individual scaled scores for the 6th grade Ohio reading proficiency test. Scores were provided for each of the years 1998 through 2003, inclusive. In accordance with federal and state privacy regulations, the office, prior to releasing the data used in this study, removed all information from the student achievement records that might enable the identification, either directly or by inferential means, of individual student data.

3.8 Missing data

Students absent due to excused or unexcused absence from both the regular administration and the make-up administration of the test are not included in the analysis. According the Director of Research, Planning and Assessment, missing records for these ‘no-shows’ amount to less than 1% in any year (Peters, 2003). For student with records on hand, it is anticipated that only a very small proportion of student records will be incomplete (e.g., missing reading score, missing school identifier). Since the study is incorporating population rather than sample data, the impact of a small amount of missing data is expected to be accordingly small. Missing data will be reported and evaluated along with the descriptive data at the beginning of the presentation of results, in chapter four, below. Implications for the study design and interpretation of the equity analyses will be addressed at that time.
3.9 Research questions

The study addresses three research questions. The questions are structured to evaluate horizontal equity in a given study district based on the logic of a general school accountability policy framework in which a district would be expected to demonstrate overall improvement in student scores while simultaneously reducing variability among those scores, particularly through improvements at the low end of the distribution.

1. Have reading achievement scores improved across the district overall?
2. Has the horizontal equity among student reading scores improved?
   a. With respect to point measures of horizontal equity?
   b. With respect to variance-based measures of horizontal equity?
   c. With respect to distributional measures of horizontal equity?
3. Has within- and between-schools student achievement equity improved?
   a. Has horizontal equity improved between schools?
   b. Has horizontal equity improved within schools?
   c. Have within- and between-schools equity converged?

3.10 Analytical method

The statistical analysis proceeds in four steps: a review of descriptive statistics followed by three separate analyses of the three research questions. In the first step, a review of descriptive statistics provides a general overview of the patterns of student performance throughout the study period both by grouping variable and overall. The descriptive report begins with a summary of missing data and corrective action or caveats, as warranted, for analysis or interpretation. The descriptive report continues
then with a summary of the number of students \( (n) \), mean, and standard deviation.

Supplementing the descriptive statistics, and of particular interest to the first research question, pass rates are computed based on the percentage of students scoring at or above the reading test cut score. The descriptive statistics offer a preliminary indication of the degree of variability within the score distributions and are also suggestive of the degree of consonance between trends in the pass rate, mean score, and score variability from year to year.

Following the discussion of descriptive statistics, the analysis turns to the research questions. Research Question 1 asks whether performance increased in the district overall during the study period. Reading proficiency rates offer the conventional method of describing student achievement on the proficiency test. A more detailed examination of the statistical and substantive significance of observed differences in average reading scores across years follows, including results of the omnibus \( F \) test from a one-way analysis of variance (ANOVA) of reading scale scores by year, appropriate follow-up tests comparing performance in the base year to each subsequent year, and analysis of linear trend across the study period.

Research Question 2 asks whether reading scores across the district have tended to converge over the study period. Results of the point measures of horizontal equity (range, restricted range, federal range ratio), variance-based measures of horizontal equity (standard deviation, coefficient of variation, standard deviation of logarithms), and distributional measures of horizontal equity (quintile analysis, Gini coefficient, Atkinson index, Theil statistic) inform Research Question 2.
Research Question 3 asks whether reading scale scores have tended to converge within and between schools, that is, whether horizontal equity has improved both within and between schools. This step utilizes decompositions of the Atkinson and Theil measures into within- and between-groups components, where the grouping variable of interest is schools. The decompositions allow direct comparisons of the relative share of overall inequality that is due to inequality within schools as compared with inequality between schools. Examining the information across the six years of the study period permits analysis of whether the shares have converged over time and whether the degree of horizontal equity in each component has changed.

3.11 Evaluation of Statistical Hypotheses

This section defines the statistical hypotheses associated with each research question and the methods employed by the analysis to evaluate those statistical hypotheses. The discussion follows the order of the three research questions.

3.11.1 Research Question 1

The null hypothesis for the first research question is that no change has occurred in overall student performance across the study period. This hypothesis is operationalized in two ways. With respect to pass rates, the null hypothesis is interpreted to mean that the district pass rate is equal for every year in the six year period, 1998-2003. State and federal accountability policy treats the pass rate as an error-free measure of achievement, implying that any observed differences are ipso facto evidence of differences in the distribution of achievement. E.g., the Ohio accountability standards in
place while the current data were being collected stipulated that districts demonstrate that
at least 75% of students pass the sixth grade reading test – with no allowance or
adjustment for error in the measurement of the reading scores or the nominal pass rate
(General Assembly of the State of Ohio, Am. Sub. S.B. 55).

As discussed earlier, evidence from the Ohio Department of Education suggests
that the test scores do include error variance, as represent by the standard error of
measure, or SEM (Ohio Department of Education, 2004). A more formal statistical
approach to the question of whether average performance has increased significantly over
the period is through the application of analysis of variance, or ANOVA, a technique that
estimates differences between years based on the observed variance in scores within and
between years. The null hypothesis for the omnibus test is:

\[ H_0 : \mu_{1998} = \mu_{1999} = \mu_{2000} = \mu_{2001} = \mu_{2002} = \mu_{2003} \]

The alternative hypothesis is that at least one of the means is different. If the
omnibus test is significant, we follow up with two sets of post hoc tests. In the first of
these we determine whether the results for the base year differ from any of the
subsequent years. Although we are interested in a directional outcome, i.e., whether
results for years subsequent to the base year are greater than the base year, the possibility
exists that, to the extent that they differ at all on average, student scores may be either
higher or lower, which calls for the use of a two-tailed test. For the pairwise tests, the
null hypothesis for each comparison is that the means are drawn from the same
underlying population with the same population mean, \( \mu \) and that, therefore, the
observed mean in the years 1999 through 2002 are not statistically significantly different
from the 1998 mean reading scale score. The alternative hypothesis for each comparison is that the observed means are drawn from populations with different means.

The evaluation of multiple hypotheses from a common dataset, such as is the present case, is associated with cumulative increases in Type I error, and hence to an increased probability of finding at least one statistically significant result by chance alone (Keppel, 1991, p. 164). The Dunnett test (Dunnett, 1955) controls cumulative, or familywise, error for all possible pairwise comparisons against a single comparison group. The technique as followed here is detailed in Keppel (1991, pp. 176-177). While the Dunnett test helps ensure against inflation of alpha, Keppel cautions that the use of any alpha-adjusting technique for multiple comparisons, including the Dunnett test, opens the opposite risk of being overly conservative (p. 177). We follow Keppel suggestion of evaluating the test results at three levels: significant (p<.025), not significant (p>.10), and not conclusive (.025<p<.10).

The second follow-up to the omnibus ANOVA is the test of linear trend (Keppel, 1991, pp. 142-147), i.e., whether the form of the relationship among reading score means conforms to a straight line of the form:

$$\bar{Y}_i' = \bar{Y}_T + bc_i$$

where $\bar{Y}_i'$ = the mean for the $i$-th year

$\bar{Y}_T$ = the grand mean for all years

$b$ = slope of the line

$c_i$ = the linear trend coefficient for the $i$-th year
The null hypothesis is that the slope, $b$, is equal to zero, i.e., that no linear trend is present. The alternative hypothesis is that the slope is not equal to zero and that, therefore, the annual reading scale score means exhibit a linear trend. Other types of trends are theoretically possible, including quadratic, cubic, etc. A problem arises with respect to interpretability of such forms, however. A cubic trend is a wave pattern of alternating peaks and troughs. From an achievement or equity perspective, such a pattern would be difficult to interpret in the conventional sense of scores ‘trending’ up or down. If the linear trend fails to account for the majority of the mean square term associated with years, the analysis will proceed to exploration of higher order trends. In light of the foregoing discussion, results of nonlinear trend will be interpreted with caution.

The analysis of variance is designed to function under certain conditions, stipulated in five assumptions: the dependent variable is measured at the interval or ratio level, observations are independent, selection and assignment of students has been randomized, scores are normally distributed, and score variances are equal. The assumptions are addressed in detail and corrective action recommended where indicated.

3.11.2 Research Question 2

Research Question 2 asks whether reading scores across the district have tended to converge over the study period. The null hypothesis for the second set of research hypotheses is that no net change in equity has occurred. The alternative hypothesis is that each of the measures indicates a shift toward greater or lesser equity. As discussed in the literature review, one of the difficulties of inequality analysis is the absence of clear benchmarks for determining either ‘statistically significant,’ ‘meaningful,’ or
‘substantial’ differences between distributions. If a positive change is observed does that change carry any practical significance? Following the computational characteristics of each of the measures, observed movement toward equality is taken to be positive, and vice versa, without reference to the qualitative magnitude of the observed changes.

The distributional measures of horizontal equity, the Gini coefficient, Atkinson index and Theil statistic, have been subjected to the widest application in the equity research literature and are generally viewed as the ‘standard’ measures of equity (see, e.g., Sen, 1997). For as much work as has been conducted with these measures, however, the literature does not deal with the question of statistical or substantive change. The literature treats these distributional equity statistics as error-free measures of equality and any changes over time are generally viewed as ‘significant’ (see for example Lynch, 2003).

In the context of achievement equity, the measures of horizontal equity are applied to student achievement scores that have a known error variance (Ohio Department of Education, 2004). This test score variability suggests a possible approach to the question of statistical significance of changes in equity values. The standard error of the mean of the sixth grade reading scale scores indicates the range within which the observed mean might vary for a given group of students if the test were to be readministered an infinite number of times. Shifting the distribution of observed scores to the left by 1.96 standard errors and to the right by 1.96 standard errors and recomputing the distributional equity statistics for each of these two shifted distributions offers a set of estimates of the upper and lower bounds of a 95% confidence interval separately computed for each distributional equity statistic.
Extending this analysis permits systematic examination and evaluation of trend following ANOVA trend analysis techniques. Since each equity statistic, noted here as $e$, representing any of the three distributional measures of horizontal equity, is computed for the observed distribution of student scores each year, the sum of squares associated with within-year variation in $e$ may be computed indirectly. First, adding to each observed student score $+/-1.96$ times the standard error of the reading scale score mean, $SE_{mean}$, for each year provides an estimate of the range of variation of the underlying population distribution of the reading scale scores. Recomputing $e$ on these transformed data provides the corresponding upper and lower bounds of the 95% confidence interval for $e$ for a given year. The sum of squares associated with within-year variation in $e$ is then isolated by algebraic manipulation of the formula for the confidence interval, given $n$, which is known for each year, and the standard error of $e$, $SE_{E}$, which is computed from the upper and lower bounds of $e$. The $F$ statistic is then computed in the usual way, following Keppel (1991), and evaluated at the alpha=0.05 level of confidence. The null hypothesis is that no linear trend is present. The alternative hypothesis is that the annual values for $e$ conform to a linear trend.

To summarize, Research Question 2 addresses the question of whether horizontal equity in the district has changed over the research period. The null hypothesis of no change is evaluated in three ways: a description of the results of the point-based-, variation-based-, and distributional measures of equity, a statistical test of equality conducted for each of the three distributional measures of horizontal equity, and a statistical test of linear trend conducted for each of the three distributional measures of horizontal equity.
3.11.3 Research Question 3

Research Question 3 asks whether reading scale scores have tended to converge within and between schools, that is, whether horizontal equity has improved both within and between schools. The null hypothesis associated with the third research question is that no net change has occurred either between or within schools from 1998 through 2003. Descriptive statistics, ANOVA, and within- and between-schools decompositions of the Atkinson index and Theil statistic enable the evaluation of the null hypothesis.

The following section describes the equity measures in the order presented in the analysis. For each of the measures, the presentation includes the mathematical formulation, a brief description of the computational characteristics.

3.12 Computational characteristics of the equity measures

As described under the Analytical Method section, above, the approach to evaluating the research questions incorporates the results of ten horizontal equity measures. The question of whether equity has improved is, as discussed in chapter two, above, a question of whether equity has improved in relation to some valued condition or conditions. These value relationships become clear when examining the computational characteristics of the measures. In order to explicitly articulate these value relationships, this section presents each of the measures, providing the mathematical formulation and a description of the salient computational characteristics of each. Understanding these characteristics illuminates the mechanics of each measure and, of particular importance, the effects that different changes within a distribution will have on a particular measure.
This information provides the basis for interpreting the results observed in chapter four, below, and for evaluating the statistical hypotheses.

The measures are grouped and discussed here in the order presented in the Analytical Method discussion, above. In order, these include: a) Point measures (range, restricted range, federal range ratio); b) Variance-based measures (standard deviation, standard deviation of logarithms, coefficient of variation); and, c) Distributional measures (Gini coefficient, Atkinson index, Theil statistic).

3.12.1 Point measures of horizontal equity

This section provides an introduction to and overview of each of the three point-measures of horizontal equity. The discussion proceeds with a presentation of the computational method for the statistic followed by a discussion of the computational characteristics and implicit value judgments that must be born in mind when interpreting the results of the statistic.

3.12.1.1 Range

The range is concerned with distance between the lowest and highest score in a distribution. The range statistic, \( R \), is computed as:

\[
R = (i_{\text{max}} - i_{\text{min}})
\]  (3.1)

where:

\( i = i^{th} \) score in the distribution of scores

\( i_{\text{max}} = \) maximum score value

\( i_{\text{min}} = \) minimum score value
The range reaches a maximum when at least one student reaches a perfect score while another student fails to score a single point. A reduction in range is taken to imply a reduction in inequity. The range is sensitive only to changes in the absolute distance between the maximum and minimum scores in a distribution. The range is not sensitive to changes within a distribution so long as the distance between the minimum and maximum scores is held constant. The emphasis on the extreme values may be a source of volatility from year to year. The lack of attention to the remainder of the distribution sharply limits the amount of information available for drawing conclusions regarding overall change.

3.12.1.2 Restricted range

The restricted range, denoted $R_{restricted}$, is similar to $R$ except that, instead of focusing on the minimum and maximum scores, it measures the distance in a set of rank-ordered scores of the 5th and 95th percentiles such that:

$$R_{restricted} = (x_{95th} - x_{5th})$$

(3.2)

where:

$i_{5th} = \text{score of the student at the 5th percentile}$

$i_{95th} = \text{score of the student at the 95th percentile}$

$R_{restricted}$ reaches a maximum when at least five percent of the students reach a perfect score while at least another five percent of students fail to score a single point. $R$ and $R_{restricted}$ are quite similar, the notable difference being the potentially reduced
volatility in the latter resulting from the focus on points somewhat removed from the extremes.

### 3.12.1.3 Federal range ratio

The federal range ratio, denoted as $F$, is the ratio of the restricted range, $R_{\text{restricted}}$, to the 5\textsuperscript{th} percentile, as follows:

$$ F = \frac{R_{\text{restricted}}}{x_{5\text{th}}} $$  \hspace{1cm} (3.3)

By relating the two values on the right, $F$ yields a score-independent representation of the information contained in the restricted range, enabling differentiation between equivalent ranges at different levels. For example, if all scores are increased by a fixed increment, say five points, the numerator will remain constant whereas the denominator will increase and $F$ will decline, indicating a more equitable condition. However, it should also be noted that such a rightward shift of all scores is indistinguishable from a rightward shift of only the 5\textsuperscript{th} and 95\textsuperscript{th} percentile scores.

As a result of the 5\textsuperscript{th} percentile being in the denominator, a decline of a given amount in the 95\textsuperscript{th} percentile will have less effect on $F$ than an equivalent increase in the 5\textsuperscript{th} percentile, ceteris paribus. This disproportionality increases as the numerator decreases. This characteristic suggests an implicitly greater value for improving equity by raising scores from the bottom up rather than by lowering scores from the top down.

The $R$, $R_{\text{restricted}}$, and $F$ are all scale dependent measures. That is, comparisons between tests of differing scales or between measures of entirely different scales, say income versus achievement, must be undertaken with caution because differences may
accrue due merely to scale differentials rather than material differences in conditions. In this regard, a math test with a scale of 0-275 and a reading test with a scale of 0-350 must be treated independently, with comparisons made among scores on the same test rather than between scores on the different tests.

3.12.1.4 Summary of the point measures of horizontal equity

The range, restricted range, and federal range ratio, provide basic but limited information about the degree of inequality in a score distribution. The restricted range improves on the range by shifting focus away from the extremes to look instead at the 5th and 95th percentile scores. The federal range ratio eliminates the unit of measurement from the analysis by examining the ratio of the restricted range to the 5th percentile score. The latter method permits an examination of inequity in the context of the overall position of the distribution. The point measures are, as described by Sen (1973; 1997), of little practical value as a result of the fact that these measures essentially ignore all but two points in any given distribution. In spite of this inherent weakness, and their disappearance from the income equity literature, the measures are included here as a result of their continued appearance in the educational equity measurement literature. Based on these computational characteristics and the results in chapter four, below, the question of the utility of these measures will be revisited in the discussion in chapter five.

3.12.2 Variance-based measures of horizontal equity

This section provides an introduction to and overview of each of the three variance-based measures of horizontal equity used in the analysis. The discussion
proceeds with a presentation of the computational method for the statistic followed by a
discussion of the computational characteristics and implicit value judgments that must be
born in mind when interpreting the results of the statistic.

3.12.2.1 Standard deviation

The standard deviation, denoted $S$, represents the square root of the mean squared
distance of scores from the mean:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}} \quad (3.4)$$

where:

$x_i$ = score of the $i^{th}$ student

$\mu$ = the arithmetic population mean

$N$ = number of students in the population

A decrease in the standard deviation implies a decrease in the overall dispersion
of scores and therefore a decrease in inequity. Reducing a score that is above the mean,
as long as the reduction is less than two times the absolute distance to the mean, leads to
lower variance. Likewise, increases a score that is below the mean, as long as the
increase is less than two times the absolute distance to the mean, leads to lower variance.

Due to the squaring of distances in the numerator, a one point change in a score in
a tail yields a greater change in variance than does a corresponding one point change
nearer the mean, implying greater value for changes among extreme scores. Using the
mean as a reference point, it is also important to note that a one point shift in any score
toward the mean reduces the distance of all other scores to the mean by $1/n$ and the
distance of the changed score by 1-(1/n), creating an impression of general improvement across scores when only one score has in fact changed.

Shifting the entire distribution right or left by a fixed increment leads to no change in the variance (i.e., absolute variation is held constant), implying that a given amount of variance is associated with a comparable degree of inequity regardless of whether the scores in question are generally very low or very high. Shifting scores rightward by means of a fixed multiple (i.e., multiplying each score by a constant, c), by contrast, leads to an increase in variance (i.e., relative variation is held constant) because the spread around the mean is increased.

The computational characteristics imply that it is increasingly desirable to raise a student’s scores the further below the mean that student’s score falls. This must be balanced, however, against the corresponding merit this measure gives to lowering the scores of students at the high end of the distribution.

A final caveat on $S$ is its reliance on the mean, a characteristic that produces a compensatory tendency to balance increases in relative variation against decreases in the mean. This suggests that greater relative variation is increasingly tolerable as scores decline overall. For more on the standard deviation as a measure of equity, the reader is referred to Sen (1997).

**3.12.2 Coefficient of variation**

The coefficient of variation, denoted here as $C$, is quotient of the standard deviation divided by the mean, such that:
\[ C = \frac{S}{\mu} \]  

(3.5)

Similar to \( S \), reductions in \( C \) imply reductions in inequity. \( C \) has a lower bound of zero, representing perfect equity. The upper bound for \( C \) must be determined for each case, based on the hypothetical situation in which all but one of the students score at the test’s minimum obtainable score, and the one remaining student achieves the test’s maximum obtainable score. From an interpretive perspective, this implies that the most inequitable situation would be one in which only one student obtains the academic benefits measured by the test.

Like \( S \), \( C \) is sensitive to changes in relative variation (i.e., equal percentage increases across scores) and insensitive to changes in absolute variation (i.e., equal unit increases across scores). Unlike \( S \), however, \( C \) is independent of the mean, a characteristic that enables \( C \) to overcome \( S \)’s tendency to trade off increases in relative variation against decreases in the mean.

Equal additions across all scores leads to a reduction in relative variation and a decrease in the coefficient. Equal percentage additions across all scores yields an increase in absolute variation but no change in relative variation and no change in the coefficient. The effect on \( C \) of a given change in a score depends on the distance of that score from the mean; the greater the distance that score is from the mean, the greater will be the effect on \( C \). Offsetting changes such that one score gains exactly the amount lost by another score higher in the distribution, so long as the offset does not result in a change in the ranking of the two scores, yields an equal effect on the coefficient irrespective of the positions of the two scores.
The computational characteristics of $C$ overall are quite similar to $S$, with the exception that whereas $S$ implies that increases in relative variation can be traded off against decreases in mean performance, $C$ implies that increases in variation are undesirable irrespective of changes in the mean.

### 3.12.2.3 Standard deviation of logarithms

The standard deviation of logarithms, denoted $SDL$, represents a logarithmic transformation of the standard deviation. The transformation may be achieved by taking any base. Following Atkinson (1970) and Sen (1997), the natural logarithm represents the most commonly used transformation in the inequality literature:

\[
SDL = \sqrt{\frac{\sum_{i=1}^{n} (\ln x_i - \ln x)^2}{N}}
\] (3.6)

As in $S$, decreasing values of $SDL$ imply greater equity, to a minimum of zero. Four computational characteristics differentiate $SDL$ from $S$. First, the logarithmic transformation produces increased sensitivity to changes at the lower end of the distribution. Given two scores equidistant above and below the mean of a uniform distribution, a unit shift toward the mean by either score will reduce $SDL$, but the effect will be disproportionately greater for the score below the mean. Second, the logarithmic transformation reduces the deviation, thereby muting overall inequality. Third, the logarithmic transformation eliminates effects relating to the unit of measure. For more details on the deviation of logarithms as a measure of equity, the reader is referred to Sen (1997).
From a policy perspective, the first computational characteristic is of $SDL$ particular relevance. In essence, this characteristic permits measurement of inequity under the assumption that increases in low scores is of substantially greater value to overall equity than lowering high scores. Among the three variance-based measures of horizontal equity, this explicit value assertion is unique to the $SDL$.

3.12.2.4 Summary of variance-based measures of inequality

The three variance-based measures reviewed here offer more comprehensive approaches than those taken with any of the three point-measures of horizontal equity, discussed above. Comparing all scores to the mean, rather than taking only two selected points in the distribution, as in the descriptive measures, offers a more comprehensive strategy for the measurement of inequality. This approach also offers a caveat in the sense that the researcher must ask whether the mean is a relevant benchmark. The coefficient of variation, $C$, and the standard deviation of logarithms, $SDL$, both improve on $S$ through the elimination of effects associated with measurement unit. Going one step further, $SDL$, offers the means of prioritizing changes at the low end of the distribution over changes in the upper end of the distribution. All three measures employ a squaring of differences, whether from the mean or the log transformed mean, calling into question whether this represents a specific rationale or merely an interesting byproduct of employing statistical measures that were originally developed for other purposes. While this procedure increases sensitivity to changes among extreme scores, neither the income inequality literature nor the educational inequality literature offer theoretical justifications for this particular mathematical choice. Also, these literatures do not discuss the policy
ramifications of offering conclusions with regard to results based on this aspect of these measures. In contrast, the SDL’s greater sensitivity to changes at the lower end of the distribution has been widely discussed, from Dalton’s (1920) seminal work to current methodological guides such as Sen (1973; 1997) and Berne & Steiffel (1984). The discussion in chapter six reexamines all of these issues based on the computational characteristics presented here and the empirical results presented in chapter four.

### 3.12.3 Distributional measures of horizontal equity

This section provides an introduction to and overview of each of the three distributional measures of horizontal equity used in the analysis. The discussion proceeds with a presentation of the computational method for the statistic followed by a discussion of the computational characteristics and implicit value judgments that must be born in mind when interpreting the results of the statistic.

#### 3.12.3.1 Quintile analysis

A useful overview of changes in a score distribution is the quintile analysis (Lynch, 2003). The information gained by this descriptive approach is analogous to the review of simple descriptive statistics researchers employ before undertaking statistical tests; the quintile analysis offers a glimpse of the structure of the distributions permitting the researcher to gain insight into the general patterns in advance of the computation of the distributional measures.

For the quintile analysis, student scores are ranked from lowest to highest. Students are then divided into five equal groups, or quintiles, $Q1$ (lowest 20% of
students) through $Q_5$ (highest 20% of students). If the number of students does not permit equal grouping, the number of students is balanced among the quintiles. For example, if there are 101 students, $Q_1$ will have the lowest 20, $Q_2$ the next 20, $Q_3$ will have the next highest 21 students, $Q_4$ will have 20, and $Q_5$ will have the highest performing 20 students. The numbers of students used in this analysis will render such quintile balancing trivial.

Three sets of statistics are computed on the quintile data. First, the quintile means represent the central tendency of each quintile. A caveat to these results is the fact that $Q_1$ will tend to be negatively skewed and $Q_5$ will tend to be positively skewed due to outliers. The second method of understanding the changes among and between quintiles from year to year is the calculation of the percentage change for each quintile from the base year, i.e., 1998 in the present example. This offers perspective on whether the pattern of change of one quintile over time is different over time from the pattern of change in another quintile over the same period.

The third method of employing the quintile data is computation of quintile shares within each year. A quintile’s share of achievement scores for a given year is the sum of scores for students in the quintile divided by the sum of scores for all students for that year. If scores are equally distributed, the share of scores in each quintile will equal the share of students in that quintile: 20%. Since, in practice, students in $Q_1$ will have lower scores than students in $Q_5$, we expect $Q_1$ to have less than 20% of the scores and $Q_5$ to have more than 20% of the scores. An analytical question is whether the pattern of shares from year to year suggests convergence or divergence. Share increases in $Q_1$ with corresponding share stability or decrease in $Q_5$ will tend to signal improvement in equity.
Understanding these descriptive data will be suggestive of the results expected in the distributional measures and may provide useful information for the interpretation of those measures. The reader is referred to Lynch (2003) for a detailed example.

3.12.3.2 Gini coefficient

The Gini coefficient, denoted $G$, is an algebraic derivation of the ratio of the areas below a Lorenz curve (Lorenz, 1905; Gini, 1912). Two distributions are represented in Figure 3.1, below. The first, shown as the diagonal line, represents the hypothetical case in which all students achieve the same score. In that case, Lorenz (1912) demonstrated that, for any case in which all persons share equal amounts of income (or, in our case, achievement scores), the cumulative distribution of income (or scores) will correspond exactly to the cumulative distribution of persons (or students) in such a way that the cumulative distributions, when plotted against one another, will yield a diagonal line, as that shown in Figure 3.1, below. By contrast, for any unequal distribution of income (or scores), and when the individual incomes (or scores) are arranged in ascending order, the resulting plot will always fall on a curve below the diagonal of equal distribution, as illustrated by the curved line in Figure 3.1, below.
Conceptually, the Gini coefficient relates the area between the diagonal line of equality and the curve of observed inequality, the area marked $A$ in Figure 3.1, to the total area under the diagonal, such that $G = A / (A + B)$. Equation 3.6, below, provides the computational formulation of the coefficient:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \mu}$$  \hspace{1cm} (3.7)

The notation in the numerator reflects the sum of differences in scores between all possible $(i,j)$ pairs of students in the population. As implied by the illustration in Figure 3.1, above, values for $G$ range from 0.0, representing perfect equality, to 1.0, representing perfect inequality (i.e., 100% of total scores obtained by single student).

The Gini coefficient is a distributional measure of horizontal equity because it takes into account the positions of all scores in the distribution with respect to each other and not merely with respect to a single point, as in $S$ or $C$, which use only the mean. The mean is still a relevant benchmark in $G$, however, to the extent that shifting a score
toward the mean yields a reduction in the sum of paired differences, i.e., the numerator of the coefficient, and also an increase in the denominator, i.e., the mean. The effect on the numerator increases with the distance from the mean at which the shift occurs.

These computational characteristics imply several value judgments. First, raising the scores of students below the mean is more desirable, though only slightly, than lowering scores of students who are above the mean. This effect increases with increased distance from the mean. Although the same is true for scores above the mean, raising the low score is always more desirable than lowering an equivalently high score. As with $C$ and $SDL$, equal additions to all scores, i.e., shifting the entire distribution rightward, reduces inequity. As with $C$ and $SDL$, equal proportional increases in all scores is equity neutral.

3.12.3.3 Theil statistic

The Theil statistic, $T$, is derived from the physical concept of entropy, the tendency of closed systems to tend toward disorder (Theil, 1967). Theil applied the concept first to information theory and later to income inequality (e.g., Sen, 1997). Theil’s equation provides the means by which to measure the degree to which each person has obtained a proportional share of income (or, in our case, achievement), as follows:
The reader may object to what appears at first to be an undersimplified equation, but the particular form of equation 3.8 lends two important benefits. First, further simplification, see e.g., Berne & Steifel (1984), Coulter (1989), Sen (1997), tends to obscures the intuitive simplicity of the formula’s integration of score share (represented in both the first quotient and the numerator in the second quotient) and population share (represented in the denominator in the second quotient). Second, equation 3.8 reveals how the logarithmic function staggers the results of the second quotient, emphasizing inequalities at the low end of the distribution. For more on the intuitive interpretation of T, the reader is referred to Conceição & Ferreira (2000).

As student scores become more evenly spread, T approaches zero. T reaches its theoretical maximum when only one student obtains a scores greater than zero. Generally, shifting any score toward the mean yields a reduction in T and vice versa. Owing to the logarithmic manipulation similar to that in SDL, above, the effect on T of shifting a score toward the mean is greater for a score below the mean than for a score above the mean, ceteris paribus. This effect is stronger than in G.

Raising scores below the mean and lowering scores above the mean are both effective strategies for reducing inequality, however raising scores below the mean is exponentially more desirable than lowering scores above the mean, an effect that
increases with increased distance from the mean. A corollary to this characteristic is that, while increasing a score above the mean is undesirable, decreasing a score that is below the mean is exponentially less desirable. Equal additions to all scores, i.e., shifting the entire distribution rightward, reduces inequity. The effect of equal proportional increases in all scores is equity neutral.

3.12.3.4 Atkinson index

The third member in the set of distributional measures of horizontal equity is the Atkinson index, denoted here, \( A \). The structural design of the index is based on a social welfare function proposed by Atkinson (1970). From the perspective of inequality measurement, however, the index has several very useful properties. For example, the index shares some characteristics in common with \( T \) by virtue of providing an alternative approach to the weighted sum of proportional scores method, as follows:

\[
A = \frac{\sum_{i=1}^{N} \left( \frac{1}{N} x_i^{1-\varepsilon} \right)^{(1-\varepsilon)}}{\mu} \quad \text{for } \varepsilon > 0 \text{ and } \varepsilon \neq 1
\]

(3.9)

Where:

\( \varepsilon \) = a constant selected by the analyst representing the degree of inequality aversion

As with \( G \), values for \( A \) vary in the range \([0, 1]\). Unlike \( G \), however, the scale is reversed, such that perfect equity is demonstrated when \( A = 1 \). This can be reversed simply by subtracting the index value from unity, as is often seen in the literature, e.g.,
Atkinson (1970); Sen (1973); Berne & Steifel (1984), etc., but this interpretive convenience is sacrificed here in favor of a substantial computational simplification in the factorial decomposition of the index, discussed later in this chapter.

The idiosyncratic element of \( A \) is the use of the inequality aversion coefficient, \( \varepsilon \). The effect of this coefficient is that, as the value of \( \varepsilon \) increases, so too does the sensitivity of the index to changes in scores below the mean. The term idiosyncratic is used here because the coefficient is an entirely arbitrary variable selected by the researcher. Atkinson, in his own work, suggested values of \( \varepsilon \) of 0.5, 1.0 (for which there is a separate, special formulation outside the scope of the present study), and 2.0. The school finance literature includes examples of researchers using values of \( \varepsilon \) as large as 30 (Berne & Stiefel, 1984). There has been as yet, however, no clear theoretical justification for the coefficient other than that it tends to produce certain properties desired by inequity researchers. As the literature has matured, researchers have adopted \( \varepsilon = 0.5 \) as a conventional value for the inequality aversion coefficient. For more on this, the reader is referred to Atkinson (1970), and, Sen (1997).

As with \( T \), any movement toward the mean by score that is below the mean will yield a disproportionately greater improvement in \( A \) than will an otherwise equally situated score above the mean. This effect, as just discussed, increases as \( \varepsilon \) increases. Otherwise, the characteristics of \( A \) are consistent with those of \( G \) and \( T \), above.

### 3.12.3.5 Summary of the distributional measures of horizontal equity

The distributional measures of horizontal equity provide three alternative approaches to incorporating consideration of every score in the distribution, the
proportion of each score to the total, and the proportionate share each member should have if all scores were distributed equally, whether the members are individual students, schools, or other unit of analysis. All three measures place some degree of greater emphasis on changes among scores at the lower end of the distribution. In this regard, $A$ is unique insofar as the researcher may control the level of this sensitivity. Two of the measures, $G$ and $A$, vary in the range of $[0, 1]$, an attractive characteristic from an intuitive perspective and, in the case of $G$, a characteristic that enables a straightforward comparison over time in terms of measuring progress toward greater equity.

3.12.4 Decomposable measures of horizontal equity

The measures presented in equations 3.1 through 3.9 address the hypotheses associated with Research Question 1. As discussed, each measure approaches the question in a slightly different way, each being sensitive to particular types of differences among distributions. Research Question 2 requires that we go a step further and break down the total observed inequality into within- and between-schools contributions to the total, observed inequity. As discussed in the literature review in chapter 2, above, the problem of the decomposability of the distributional inequality measures has challenged researchers for over thirty years. Presently, two measures are available, the Theil statistic and Atkinson index. The Theil measure, $T$, is chief among the two because it has been shown to be additively decomposable, e.g., Conceição & Ferreira (2000), that is, separable into within- and between-groups components such that

$$T = T_b + T_w$$  \hspace{1cm} (3.10)
The Atkinson index, while decomposable, requires a factorial approach, such that:

\[ A = A_b A_w \]  \hspace{1cm} (3.11)

The remainder of this section will provide the mathematical computations for these decompositions. The computational characteristics and interpretive considerations are consistent, in both cases, with the parent formulas discussed above, so no further discussion is offered here with respect to the particular characteristics of the decompositions.

3.12.4.1 Decomposition of the Theil statistic

Working from equation 3.8, above, the between-groups portion of total inequality can be expressed by replacing individual students with individual schools. The decomposition takes the following form for the between-groups component:

\[
T_b = \sum_{j=1}^{J} \frac{x_j}{n} \ln \left( \frac{x_j}{\sum_{i=1}^{n} x_i} \right) \left( \frac{n_j}{N} \right)
\]  \hspace{1cm} (3.12)

Where:

\[ J = \text{number of schools} \]
\[ j = j^{th} \text{ school} \]
\[ x_j = \text{sum of student scores in the } j^{th} \text{ school} \]
\[ n_j = \text{number of students in the } j^{th} \text{ school} \]
Completing the decomposition, the within groups component, $T_w$ is computed as:

$$T_w = \sum_{j=1}^{J} \left( \frac{x_j}{\sum_{i=1}^{n_j} x_i} T_j \right)$$

(3.13)

Where:

- $j$ = the $j^{th}$ school
- $n_j$ = number of students in school $j$
- $T_j$ = Theil measure for the $j^{th}$ school as individually computed via eq. 3.8 for each school

These individual subtotals of the Theil statistic follow the same characteristics as govern the general formula, discussed above. Reviewing equations 3.8, 3.12 and 3.13, the reader may appreciate the overall continuity and, hopefully, intelligibility of the basic equation and its decomposition. The reader may also note with interest that the additive nature of the Theil decomposition can readily be extended to additional levels. For example a state might be subdivided into three levels, representing inequality between districts, between schools, and within schools. The recommendations for further research, in chapter five, below, address this possibility in greater detail.

### 3.12.4.2 Decomposition of the Atkinson index

Turning to the decomposition of the Atkinson index of inequality, $A$, Lasso de la Vega & Urrutia (2003) have demonstrated a factorial decomposition, such that:

$$A = A_w A_w$$

(3.13)
Following equation 3.9 and Lasso de la Vega & Urrutia (p. 3), the between-schools factor is computed:

$$A_b = \left[ \sum_{j=1}^{J} \left( \frac{N_j}{N} \mu_j \right) \right]^{-\frac{1}{1-\varepsilon}}$$

for $\varepsilon > 0$ and $\varepsilon \neq 1$ (3.14)

and the within-schools component is computed:

$$A_w = \left[ \sum_{j=1}^{J} \left( \frac{N_j}{N} \right)^{\varepsilon} \left( \frac{x_j}{N \sum_{i=1}^{N} x_i} \right)^{1-\varepsilon} \right]^{-\frac{1}{1-\varepsilon}}$$

$$\left[ \sum_{j=1}^{J} \left( \frac{N_j}{N} \right)^{\varepsilon} \left( \frac{x_j}{N \sum_{i=1}^{N} x_i} \right)^{1-\varepsilon} \right]^{-\frac{1}{1-\varepsilon}}$$

(3.15)

All notation is as discussed previously. A limitation that may be immediately apparent to the reader is that this factorial decomposition does not lend itself to extension beyond two levels, i.e., within and between schools or within and between districts. No strategy has yet been reported for a decomposition of $A$ that would enable, as with $T$, a single analysis of inequality incorporating between districts, between schools, and within schools elements based on the total observed inequality. For a further discussion of decomposability and related issues, the reader is referred to Atkinson, 1970; Sen, 1973; Coulter, 1989; Conceição & Ferreira (2000); Sala-i-Martin (2002); and, Lasso de la Vega & Urrutia (2003).
3.12.4.3 Summary of the decomposable measures of horizontal equity

For any set of grouped data, decomposition of the equity measures enables researchers to examine discretely the contributions to inequity made by variation between groups as opposed to variation within groups. The Atkinson index, $A$, permits this decomposition in a factorial manner, limiting the analysis to a single set of groups. In spite of this limitation, the index has the benefit of varying in the range $[0, 1]$, and of giving the researcher control over the degree of sensitivity shown to scores in the lower end of a distribution. The Theil measure, $T$, by comparison, permits multiple grouping levels by virtue of the additive nature of the decomposition. While $T$’s upper range is not intuitively simple to grasp, the computational characteristics lend much of the same interpretive qualities, such as the sensitivity to changes at the lower end of the distribution, while avoiding the arbitrariness inherent in the selection in $A$ of the coefficient of inequality aversion, $e$.

3.13 Chapter summary

This chapter opened with a restatement of the research questions, then presented the methodology for the data collection and analysis. The methodological approach is designed to examine variation in the equity of student achievement scores over the period 1998-2003 for a single, large school district, encompassing twelve elementary school buildings and approximately 700 sixth-grade students per year. Research Question 1 is examined based on the observed trend in proficiency rates and also based on a one-way ANOVA of proficiency scores by year. Research Question 2 is examined based on three sets of equity measures, including three point-based measures, three variance-based...
measures, and three distributional-measures. Research Question 3 is examined based on the factorial decomposition of the Atkinson index and the additive decomposition of the Theil statistic.

Chapter four, following, presents the results of the analysis and the evaluations of the hypotheses. Chapter five then provides a summary of the study results, both with respect to the research questions and each of the measures. Chapter six offers discussion of the study findings, conclusions drawn from the results, and recommendations for future research.
CHAPTER 4
RESEARCH FINDINGS

4.1 Introduction

This chapter presents the results of the analysis of sixth-grade reading scale scores across the study school district for the period 1998 through 2003. The ordering of results corresponds to the three research questions set forth in the introductory chapter. Before moving to the research questions, a review of descriptive statistics, including means and standard deviations, provides a score-based portrayal of district performance for each year, an indication of missing data, and a general indication of the conditions in the district with respect to the three research questions. Research Question 1 asks whether performance increased in the district overall during the study period. Reading proficiency rates offer the conventional method of describing student achievement on the proficiency test. A more detailed examination of the statistical and substantive significance of observed differences in average reading scores across years follows, including results of the omnibus $F$ test from a one-way analysis of variance (ANOVA) of reading scale scores by year, appropriate pair-wise tests comparing performance by year, and analysis of linear trend across the study period.

Research Question 2 asks whether reading scores across the district have tended to converge over the study period. Results of the point-measures of horizontal equity
(range, restricted range, federal range ratio), variance-based measures of horizontal equity (standard deviation, coefficient of variation, standard deviation of logarithms), and distributional measures of horizontal equity (quintile analysis, Gini coefficient, Atkinson index, Theil statistic) inform Research Question 2.

Research Question 3 asks whether reading scale scores have tended to converge within and between schools. Descriptive statistics by school, including the variance components of a one-way ANOVA and results of a quintile analysis of scores, offers a preliminary overview of changes in the score distributions over the study period. Decompositions of the Atkinson index and Theil statistic offer evidence for the evaluation of Research Question 3. The chapter closes with a summary of findings with respect to each of the three research questions.

4.2 Research Question 1: Has reading performance increased in the district overall?

This section is subdivided into four sections, covering the description of reading proficiency rates, score means and standard deviations, analysis of variance of scores by year, and summary of evidence for Research Question 1.

4.2.1 Reading proficiency rates

The conventional statistic used to describe district-level student proficiency is the proficiency rate. Table 4.1 presents proficiency rates for Ohio’s grade six test of reading proficiency for the study district for the years 1998 through 2003.
Table 4.1: Sixth Grade Reading Proficiency Rates, 1998-2003

<table>
<thead>
<tr>
<th></th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Proficient</td>
<td>76.2</td>
<td>74.2</td>
<td>71.8</td>
<td>80.2</td>
<td>74.5</td>
<td>81.0</td>
</tr>
</tbody>
</table>

The results indicate that, between the years 1998 and 2003, the percentage of the district’s sixth graders achieving at or above the Ohio Board of Education’s criterion for reading proficiency increased from 76.2 percent to 81.0 percent. In spite of the net gain, however, the proficiency rates in the intervening years do not appear to have fluctuated in a linear manner. Figure 4.1 offers a visual representation of these data and helps illustrate the apparent lack of linearity. The reader is advised to note the scale of the vertical axis. The narrow range of the scale aids in distinguishing the pattern among proficiency rates at the risk of exaggerating relative annual changes. The reader should understand therefore that, on a scale of 0-100 percent, the changes observed here would appear quite small by comparison.
4.2.2 Means and standard deviations

Table 4.2 presents descriptive statistics for reading scores for the district for the study years, 1998-2003. The number of student records for 1998 (n=679) is lower than for the subsequent years. According to the district’s Director of Research, Assessment and Planning, the lower number of records reflects not a smaller number of students in the district but rather incompleteness of the historic record owing to changes during the intervening years in record keeping and the database system (Peters, 2004).
<table>
<thead>
<tr>
<th>Year</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>679</td>
<td>232.2</td>
<td>27.034</td>
</tr>
<tr>
<td>1999</td>
<td>789</td>
<td>231.5</td>
<td>27.399</td>
</tr>
<tr>
<td>2000</td>
<td>843</td>
<td>229.6</td>
<td>26.419</td>
</tr>
<tr>
<td>2001</td>
<td>759</td>
<td>235.6</td>
<td>26.449</td>
</tr>
<tr>
<td>2002</td>
<td>843</td>
<td>235.7</td>
<td>30.816</td>
</tr>
<tr>
<td>2003</td>
<td>888</td>
<td>238.2</td>
<td>26.221</td>
</tr>
</tbody>
</table>

Table 4.2: Descriptive Statistics for 6th Grade Reading, 1998-2003

Are the available data for 1998 reasonably representative of the student body? It is possible that certain student subgroups are overrepresented in the missing records. To the extent that the extant records are not representative of the full class, interpretations of the results for that year might be misleading. Examining the 1998 data set in terms of representativeness of the overall student body, the proportions of females to males (49.1 to 50.9), and minority to non-minority (11.8 to 88.2) are similar to 1999. The 1998 mean (232.2) and standard deviation (27.034) appear to be consistent with the subsequent year’s results. Based on the observations made by the district administrator that the missing records are most likely random, and the evidence of similarity between 1998 and 1999 results, the remainder of the analysis is conducted on the assumption that the 1998 data represent a large and representative sample of that year’s scores.

The research question focuses on whether over the course of the study period the distributions of student scores have shifted rightward. The pattern of means in Table 4.2 suggests that the distribution of sixth-grade reading scores in this district declined from 1998 through 2000 and increased from 2001 through 2003, with a net rightward shift over the period of $238.2 - 232.2 = 6.0$ scale score points.
4.2.3 Analysis of variance

A more formal statistical approach to the question of whether average performance has increased significantly over the period is through the application of analysis of variance (ANOVA). The null hypothesis for the omnibus test is:

\[ H_0 : \mu_{1998} = \mu_{1999} = \mu_{2000} = \mu_{2001} = \mu_{2002} = \mu_{2003} \]

The alternative hypothesis is that at least one of the means is different. If the omnibus test is significant, we follow up with two sets of *post hoc* tests. In the first of these we determine whether the results for the base year differ from any of the subsequent years. Although we are interested in a directional outcome, i.e., whether results for years subsequent to the base year are greater than the base year, the possibility
exists that, to the extent that they differ at all on average, student scores may be either higher or lower, which calls for the use of a two-tailed test. For the pairwise tests, the null hypothesis for each comparison is that the means are drawn from the same underlying population with the same population mean, $\mu$ and that, therefore, the observed mean in the years 1999 through 2002 are not statistically significantly different from the 1998 mean reading scale score. The alternative hypothesis for each comparison is that the observed means are drawn from populations with different means.

As discussed in chapter three, above, the evaluation of multiple hypotheses from a common dataset, such as is the present case, is associated with cumulative increases in Type I error, and hence to an increased probability of finding at least one statistically significant result by chance alone (Keppel, 1991, p. 164). As noted, we employ the Dunnett test of pair-wise comparisons (Dunnett, 1955) with 1998 as the base year. Following Keppel, we evaluate Dunnett the test results at three levels: significant ($p<.025$), not significant ($p>.10$), and not conclusive ($0.025<p<.10$).

The second follow-up to the omnibus ANOVA is the test of linear trend (Keppel, 1991, pp. 142-147), i.e., whether the form of the relationship among reading score means conformed to a straight line of the form:
\[ \bar{Y}_i' = \bar{Y}_r + bc_i \]

where \( \bar{Y}_i' \) = the mean for the \( i \)-th year

\[ \bar{Y}_{T_i} = \text{the grand mean for all years} \]

\( b = \text{slope of the line} \)

\( c_i = \text{the linear trend coefficient for the} \ i \text{-th year} \)

The null hypothesis is that the slope, \( b \), is equal to zero, i.e., that no linear trend is present. The alternative hypothesis is that the slope is not equal to zero and that, therefore, the annual reading scale score means exhibit a linear trend.

### 4.2.3.1 Assumption checks for ANOVA

This section reviews evidence regarding whether the design satisfies the five assumptions of ANOVA (interval level measure of dependent variable, independence of observations, randomization, normality, and equality of variances). Each assumption is evaluated in turn with a judgment of whether the assumption is met. A summary at the end of this section details the status of the assumption checks and provides specific recommendations for corrective action.

### 4.2.3.2 Dependent variable measured at interval or ratio level

The first assumption, interval data, requires that the dependent variable, in this case scaled reading scores, be measured at least at the interval level. Scaled scores are linear transformation of student’s raw scores. Raw scores are measured on a continuous integer scale ranging from 0 to 38 (Ohio Department of Education, 2004), representing
equal-interval data where the interval is unity. As a linear transformation of interval data, scaled scores might be understood to be interval data. One apparent difficulty with this inference is the fact that observed intervals between student scores are not equal; as a result of the distribution of item difficulties, observed scores of 30, 31, 32 may correspond to scaled scores of 211, 215, 217, suggesting unequal intervals between scores. This appearance is, however, misleading because underlying the transformation is a continuous scale of theoretical difficulty to which the raw scores are assigned based on the observed difficulty of the items on the test (ibid). We conclude therefore that since scaled scores are measured on a continuous, equal-interval scale, the assumption is met.

4.2.3.3 Independent observations

The assumption of independence requires that, within each year and between years, each student’s score is independent of all other student scores. Independence implies that the precondition of the students is equal other than differences in reading competency (Kim, 2002). The violation of independence implies that confounding factors other than the factor of interest, i.e., student reading competency, might be present in the score data. Confounding factors might be manifest as spurious systematic variance, thereby inflating the probability of Type I error, or manifest as spurious random variance, thereby inflating the probability of Type II error. One assurance of independence is that the test is developed and administered in a secure manner, offering a discrete measure of student achievement without overt, at least, interaction between students or teachers. Student scores are evaluated across the entire district, mixing the
results of students from all sixth grade teachers in the district and thereby mitigating to some extent the teacher effects. For the decomposed equity measures, scores are evaluated school by school, a somewhat weaker aggregation. Reinforcing the grouping of data is the fact that the district was engaged in systematic curriculum planning based on the state’s model curriculum. While this will not eliminate teacher effects, teaching the curriculum model at least helps homogenize the educational experience and helps ensure that students are receiving the adequate, basic instruction necessary to pass the test (Peters, 2003). Unlike the math, science, and citizenship tests, the reading test is not merely an assessment of sixth grade content knowledge but reflects cumulative learning of reading skills. To the extent that sixth-grade teachers might have grade-specific effects, the reading test reflects skills accumulated over the previous six years plus the first three quarters of sixth grade. Finally, the district attendance zone geography tends to mix students from multiple, diverse neighborhoods, muting local socioeconomic effects (Peters, 2004). The solution to violations of independence is to shift the unit of analysis from the student to the classroom or school level. In this case, we are examining changes in mean scores at the district level, adequately ameliorating any classroom-level violations of the assumption. The impact of potential confounding due to nonindependence of observations is unknown and therefore some caution should be exercised in interpreting the results of the ANOVA. The implications of this caveat are reviewed in chapter six.
4.2.3.4 Randomization

One means of dealing with the non-independence problem is to measure the confounding variables. In this *post hoc* analysis of reading scores, however, no such data are available. Another means of handling confounding variables is randomization: the random selection and assignment of students to classrooms. This would ensure that any measurement error due to unmeasured variables would be spread across the district, emerging as random measurement error (Keppel, 1991). While students were not randomly assigned across the district, the assignment of students to schools does involve a substantial proportion of non-local attendance. That is, for capacity and logistical reasons, a student may be assigned to the nearest school or may be assigned to a school across the district. The district does not publish exact statistics on non-local student assignments, but the Director of Research, Assessment, and Planning suggested that the assignment patterns tend to break down at least some of the local homogeneity of school neighborhoods (Peters, 2004). We conclude that the assumption of randomization is not met. The implication is that the results of this study cannot be generalized outside the given district or to other years. We must also recognize the possibility that uncontrolled factors may influence student scores in certain schools or across certain subpopulations in ways that we have not anticipated. Given the infeasibility of a randomized study in a large public school system, the appropriate course of action will be to replicate the present study in other years and across other districts and across alternative decompositions in order to begin to determine the extent to which observable patterns appear to hold.
4.2.3.5 Normality of underlying population distributions

The assumption of normality, i.e., that the distribution of scores within each year
is normal, is checked by means of the Shapiro-Wilk test of normality. The confidence
level for this test is alpha=.05. Table 4.3 provides a summary of the test results.

<table>
<thead>
<tr>
<th>Year</th>
<th>Statistic</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>.986</td>
<td>676</td>
<td>.000</td>
</tr>
<tr>
<td>1999</td>
<td>.987</td>
<td>786</td>
<td>.000</td>
</tr>
<tr>
<td>2000</td>
<td>.976</td>
<td>840</td>
<td>.000</td>
</tr>
<tr>
<td>2001</td>
<td>.993</td>
<td>756</td>
<td>.001</td>
</tr>
<tr>
<td>2002</td>
<td>.990</td>
<td>840</td>
<td>.000</td>
</tr>
<tr>
<td>2003</td>
<td>.991</td>
<td>885</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 4.3: Shapiro-Wilk Test of Normality of Reading Scale Scores by Year, 1998-2003

The results in Table 4.3 suggest that the distributions of reading scale scores are
significantly non-normal in all of the study years. We conclude therefore that the
assumption of normality is violated for all groups.

The analysis of variance is generally robust with respect to non-normality if the
distribution is symmetric and the number of subjects in each group is greater than 12
(Keppel (1991, p. 97). If the distributions are asymmetric, that is, non-normal with
respect to skewness, the $F$ statistic may become inflated. In order to determine whether
the violations of normality noted in Table 4.4 are due to skewness, Table 4.4 presents the
skewness statistics for each year.
<table>
<thead>
<tr>
<th>Year</th>
<th>Statistic</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>-0.290*</td>
<td>0.094</td>
<td>-0.474</td>
<td>-0.106</td>
</tr>
<tr>
<td>1999</td>
<td>-0.189*</td>
<td>0.087</td>
<td>-0.360</td>
<td>-0.018</td>
</tr>
<tr>
<td>2000</td>
<td>0.440*</td>
<td>0.084</td>
<td>0.275</td>
<td>0.605</td>
</tr>
<tr>
<td>2001</td>
<td>0.005</td>
<td>0.089</td>
<td>-0.169</td>
<td>0.179</td>
</tr>
<tr>
<td>2002</td>
<td>0.216*</td>
<td>0.084</td>
<td>0.051</td>
<td>0.381</td>
</tr>
<tr>
<td>2003</td>
<td>-0.108</td>
<td>0.082</td>
<td>-0.269</td>
<td>0.053</td>
</tr>
</tbody>
</table>

*Note:* * significant at .05 level.

Table 4.4: Skewness Statistics for Reading Scale Scores, 1998-2003

The results suggest that the distributions in 1998, 1999, 2000, and 2002 are significantly non-normal with respect to skewness. We conclude therefore that these distributions are significantly asymmetrical. Keppel suggests that, although no direct solution is available for asymmetry a reasonable correction is to evaluate the $F$ test at a more conservative level of confidence (ibid, pp. 97-98). We will first complete our review of assumptions before proposing corrective action for these violations.

### 4.2.3.6 Equality of underlying population variances

The assumption of homoscedasticity is checked by means of the $F_{\text{max}}$ test (Keppel, pp. 100-108), which is the quotient of the largest within-group variance divided by the smallest within-group variance ($V_{\text{max}}/V_{\text{min}}$). Distributions are treated as conforming to the assumption of homoscedasticity for $F_{\text{max}}<3.0$. Results for the present case, $F_{\text{max}}=949.65/687.54=1.38$, suggest that the heterogeneity is not critical. We conclude therefore that the assumption of homoscedasticity is met.
4.2.3.7 Summary of Assumption Checks

Table 4.5 summarizes the results of the review of five assumptions underlying ANOVA. Violations are noted for three assumptions: independence, randomization, and normality. In response to the violations of independence and randomization, and the fact that no specific evidence is available regarding the possible presence of confounding variables, we will interpret the results of the tests with caution (following Stevens, 1996). In response to the violations of normality, the $F$ test is evaluated at alpha=.025 rather than the conventional alpha=.05 (following Keppel, 1991).

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Status</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interval level measure</td>
<td>No evidence of violation</td>
<td>None</td>
</tr>
<tr>
<td>Independent observations</td>
<td>No evidence of violation, but logical possibility</td>
<td>Interpret results with caution; Replicate study</td>
</tr>
<tr>
<td>Randomization</td>
<td>Evidence of violation</td>
<td>Results not generalizable; Replicate study</td>
</tr>
<tr>
<td>Normality</td>
<td>Evidence of violation</td>
<td>Possibly inflated $F$ statistic</td>
</tr>
<tr>
<td>Homoscedasticity</td>
<td>No evidence of violation</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 4.5: Summary of Assumption Checks

4.2.3.8 Omnibus test of equality of means

The results of the analysis of variance are summarized in Table 4.6. Based on the results of the One-Way ANOVA, $F(5, 4777)=11.654$, $p<.025$, we reject the null hypothesis. The results suggest that at least one of the mean reading scale scores is not equal to the others across years. The value of $r$-squared ($r$-squared=.012) suggests that 1.2% of the overall score variance is associated with annual differences in average scores. This statistic offers some insight into the question of whether the statistically significant
A statistic is substantively meaningful. Cohen (1977, pp. 284-288) defines an effect size of .01 as a ‘small’ effect and Keppel (1991, p. 67) cautiously asserts that .01 is the lower limit of a meaningful effect. The results in Table 4.6 imply therefore that although the observed differences in average scores across the six-year study period are statistically significant, those differences are substantively weak. This is borne out in an examination of the means, revealing that the maximum difference between means is the 8.6 point difference between 2000 and 2003, a difference of nearly one standard error of measure, SEM=10.7 (Ohio Department of Education, 2003).

The results warrant two sets of post hoc tests: a test of pairwise comparisons among years and a test of linear trend across years. Pairwise comparisons reveal which of the means for the six years statistically significantly differ from each other. Of particular interest here is the change in mean score with respect to the base year, 1998.

The test of linear trend reveals whether the observed differences in average scores exhibit the type of linear increase expected based on the educational reform policies described in chapters 2 and 3. The results of the pairwise and linear tests are discussed in order, below.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>R²</th>
<th>Post Study Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>43,924</td>
<td>5</td>
<td>8785</td>
<td>11.654*</td>
<td>&lt;.001</td>
<td>.012</td>
<td>&gt;.999</td>
</tr>
<tr>
<td>Error</td>
<td>3,600,925</td>
<td>4,777</td>
<td>754</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3,644,850</td>
<td>4,782</td>
<td>754</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*significant at .025 level.

Table 4.6: Summary of One-Way ANOVA of Reading Scale Scores by Year
4.2.3.9 Pairwise post hoc comparisons to base year

Table 4.7 summarizes the results of the Dunnett test of pairwise comparisons of mean reading scale scores for each of the years 1999-2003 against the base year, 1998. The results for 1999 compared with 1998 (mean difference = -.070, standard error = 1.440, $p > .025$), 2000 compared with 1998 (mean difference = -2.60, standard error = 1.419, $p > .025$), 2001 compared with 1998 (mean difference = 3.39, standard error = 1.453, $p > .025$), and 2002 compared with 1998 (mean difference = 3.54, standard error = 1.419, $p = .05$), suggest that the observed mean differences are not statistically significantly different from zero. Results of the Dunnett test of the difference in mean reading scale score between 2003 and 1998 (difference of means = 6.07, standard error = 1.402, $p < .025$), imply that the observed difference in means is statistically significantly different from zero.

Based on the results of the Dunnett test we conclude that, on average, student reading scale scores did not change either from 1998 to 1999 or from 1998 to 2000 and that scores increased significantly from 1998 to 2003. Although the results for 2001 and 2002 were not significant when evaluated at the confidence level of alpha = .025, the results fall into the ‘not conclusive’ range of .025 < $p$ < .10 and we accordingly suspend judgment as to significance.
<table>
<thead>
<tr>
<th>Year (i)</th>
<th>Year (j)</th>
<th>Difference (i-j)</th>
<th>Standard Error</th>
<th>p</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>1998</td>
<td>-0.70</td>
<td>1.44</td>
<td>0.983</td>
<td>-4.30</td>
<td>2.90</td>
</tr>
<tr>
<td>2000</td>
<td>1998</td>
<td>-2.60</td>
<td>1.419</td>
<td>0.226</td>
<td>-6.14</td>
<td>0.94</td>
</tr>
<tr>
<td>2001</td>
<td>1998</td>
<td>3.39</td>
<td>1.453</td>
<td>0.076</td>
<td>-0.24</td>
<td>7.02</td>
</tr>
<tr>
<td>2002</td>
<td>1998</td>
<td>3.54</td>
<td>1.419</td>
<td>0.050</td>
<td>0.00</td>
<td>7.09</td>
</tr>
<tr>
<td>2003</td>
<td>1998</td>
<td>6.07*</td>
<td>1.402</td>
<td>&lt;0.001*</td>
<td>2.57</td>
<td>9.57</td>
</tr>
</tbody>
</table>

*Note: * significant at .025 level.

Table 4.7: Summary of Dunnett Test of Reading Scale Score Means, 1998 vs 1999-2003

Based on the possible concern that the smaller \( n \) in 1998 might somehow limit the comparability of students scores for that year with scores from the other years, an alternative approach to the analysis of post-hoc comparisons is the Scheffé test of pairwise comparisons (Keppel, 1991). This approach provides analysis of all possible pairwise comparisons of all years with all other years in the study period. As with the Dunnett test, Scheffé controls for familywise error by adjusting the level of alpha such that the per-test critical value for \( F \) is set to \((\alpha-1)\) times the critical value for \( F \) for the omnibus test. This is analogous to dividing the nominal alpha value by \((\alpha-1)\), a conservative approach that should safely ensure control of familywise error (the reader is referred to Keppel, 1991, for computational details).

Table 4.8 presents the results of the pairwise comparisons. Note first that the results for 1998 are not dissimilar from those of 1999 in terms of statistically significant mean differences with the other years. We observe statistically significant differences between 2000 and 2001, 2000 and 2002, and we see that 2003 is significantly larger than 1998, 1999, and 2000. These results suggest that the declines observed in 1999 and 2000
were statistically not different from zero. Although the improvement in 2001 was statistically significant compared with 2000, the increase was not enough to be significant compared with 1998 or 1999, suggesting that, overall, the district still had made no net improvement by 2001. Only in 2003 do we observe statistically significant differences from 1998 and 1999. These results are consistent with the findings from the Dunnett test and does not offer any evidence suggestive of 1998 being uncharacteristic as base year.

<table>
<thead>
<tr>
<th>Year (i)</th>
<th>Year (j)</th>
<th>Difference (i-j)</th>
<th>Standard Error</th>
<th>p</th>
<th>95% Confidence Interval Lower</th>
<th>95% Confidence Interval Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>1998</td>
<td>-0.7</td>
<td>1.440</td>
<td>0.999</td>
<td>-5.49</td>
<td>4.10</td>
</tr>
<tr>
<td>2000</td>
<td>1998</td>
<td>-2.6</td>
<td>1.419</td>
<td>0.645</td>
<td>-7.32</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>1999</td>
<td>-1.9</td>
<td>1.363</td>
<td>0.856</td>
<td>-6.44</td>
<td>2.63</td>
</tr>
<tr>
<td>2001</td>
<td>1998</td>
<td>3.4</td>
<td>1.453</td>
<td>0.365</td>
<td>-1.45</td>
<td>8.23</td>
</tr>
<tr>
<td></td>
<td>1999</td>
<td>4.1</td>
<td>1.399</td>
<td>0.129</td>
<td>-0.57</td>
<td>8.74</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>6.0</td>
<td>1.376</td>
<td>0.002*</td>
<td>1.41</td>
<td>10.57</td>
</tr>
<tr>
<td>2002</td>
<td>1998</td>
<td>3.5</td>
<td>1.419</td>
<td>0.284</td>
<td>-1.18</td>
<td>8.26</td>
</tr>
<tr>
<td></td>
<td>1999</td>
<td>4.2</td>
<td>1.363</td>
<td>0.085</td>
<td>-0.29</td>
<td>8.78</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>6.1</td>
<td>1.340</td>
<td>0.001*</td>
<td>1.68</td>
<td>10.60</td>
</tr>
<tr>
<td></td>
<td>2001</td>
<td>0.2</td>
<td>1.376</td>
<td>1.000</td>
<td>-4.43</td>
<td>4.73</td>
</tr>
<tr>
<td>2003</td>
<td>1998</td>
<td>6.1</td>
<td>1.402</td>
<td>0.002*</td>
<td>1.40</td>
<td>10.74</td>
</tr>
<tr>
<td></td>
<td>1999</td>
<td>6.8</td>
<td>1.346</td>
<td>&lt;0.001*</td>
<td>2.29</td>
<td>11.25</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>8.7</td>
<td>1.323</td>
<td>&lt;0.001*</td>
<td>4.27</td>
<td>13.07</td>
</tr>
<tr>
<td></td>
<td>2001</td>
<td>2.7</td>
<td>1.360</td>
<td>0.566</td>
<td>-1.85</td>
<td>7.21</td>
</tr>
<tr>
<td></td>
<td>2002</td>
<td>2.5</td>
<td>1.323</td>
<td>0.600</td>
<td>-1.87</td>
<td>6.93</td>
</tr>
</tbody>
</table>

*Note.* *significant at .025 level.

Table 4.8: Summary of Scheffé Pairwise Test of Reading Scale Score Means, 1998-2003

**4.2.3.10 Post hoc test of linear trend**

The second follow-up to the omnibus ANOVA, above, is the test of linear trend. We examine whether the differences among means are associated with a positive, linear
trend. The null hypothesis is that no trend is present. The alternative hypothesis is that the means exhibit a linear trend. Table 4.9 presents the results of the test.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>29,481</td>
<td>1</td>
<td>29,481</td>
<td>39.109*</td>
</tr>
<tr>
<td>Error</td>
<td>3,600,925</td>
<td>4,777</td>
<td>754</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3,644,850</td>
<td>4,782</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*significant at .025 level.

Table 4.9: Summary of One-Way ANOVA Linear Trend Analysis of Reading Scale Scores Across Years, 1998-2003

Based on these results, $F(1, 4777)=39.109$, $p<.025$, we reject the null hypothesis that no linear trend is evident. The results of the test of linear trend and the plot of reading scale score means suggest that the observed means are associated with a statistically significant, positive linear trend over the study period. In terms of magnitude of the observed trend, the effect size for the linear trend component ($r$-squared=.008) is smaller than the level identified both by Cohen (1977) and Keppel (1991) as the smallest effect size of interest, suggesting that although the result of the $F$ test is significant, the magnitude of the linear component is negligible. From a public policy perspective, the fact that even a weak linear trend is evident, combined with the fact that the average score (238.2 in 2003) is well above the state’s cut score for proficiency, suggests that the results might nevertheless carry relevance, though of limited degree.

Technically, the trend analysis can be extended to higher order trends, i.e., quadratic, cubic, quartic, quintic. A critical limitation is the fact that the linear trend
accounted for the majority of the sum of squares associated with the effect, years. The remaining trends will account for the remainder, but none can be substantive in light of the weakness of the linear trend and the paucity of remaining effect variance. A second limitation is interpretability. Although the quadratic trend might be understood as a simple curve, the cubic, quartic, and quintic trends amount to undulating waves that offer little in the way of practical explanatory information other than, ‘the scores are going up and down from year to year.’ Given the limited remaining effect variance, the limited differences in mean scores for the study period, and the potential interpretive complexity, the higher order trends are not recommended.

4.2.4 Summary of evidence regarding Research Question 1

Annual reading proficiency rates reveal a net gain over the period 1998-2003 but the pattern across the intervening years appears to be volatile and does not appear to conform to a linear trend. Results of the analysis of variance suggest that the reading scale scores differ across years, that student scores in 2003 are significantly and substantively greater than in 1998, and that the annual means conform to a significant but substantively negligible linear trend. This finding is consistent with the visual examination of means in Table 4.2 and Figure 4.2, which suggested that the differences between means appeared relatively flat.

4.3 Research Question 2: Have reading scores converged across the district?

This section presents results of the descriptive, variance-based, and distributional measures of horizontal equity. Evidence is evaluated with respect to overall convergence
and with respect to differential sources of convergence, i.e., rising scores on the left side of the distribution (i.e., lower scorers getting better) as contrasted with declining scores on the right side of the distribution (i.e., higher scorers getting worse).

4.3.1 Descriptive measures of horizontal equity

Table 4.10 presents the results for the descriptive measures of horizontal equity for reading scale scores for the study district for the years 1998-2003. To the student of inferential statistics, the information in the table will likely appear somewhat curious. Being generally accustomed to seeing measures of central tendency and measures of dispersion grouped together in a single table of descriptive statistics, the descriptive measures of horizontal equity will seem to be out of context. From an equity perspective, however, we treat the types of measures in turn based on their characteristic functions. The reader will find it helpful to move back and forth between the tables for each set of equity measures, plus the descriptive statistics and figures, above.

<table>
<thead>
<tr>
<th>Year</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Restricted Range Lower</th>
<th>Restricted Range Upper</th>
<th>Federal Range</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>134</td>
<td>324</td>
<td>190</td>
<td>187</td>
<td>277</td>
<td>90</td>
<td>0.481</td>
</tr>
<tr>
<td>1999</td>
<td>147</td>
<td>319</td>
<td>172</td>
<td>183</td>
<td>274</td>
<td>91</td>
<td>0.497</td>
</tr>
<tr>
<td>2000</td>
<td>135</td>
<td>333</td>
<td>198</td>
<td>187</td>
<td>276</td>
<td>89</td>
<td>0.476</td>
</tr>
<tr>
<td>2001</td>
<td>145</td>
<td>337</td>
<td>192</td>
<td>191</td>
<td>277</td>
<td>86</td>
<td>0.450</td>
</tr>
<tr>
<td>2002</td>
<td>149</td>
<td>346</td>
<td>197</td>
<td>185</td>
<td>284</td>
<td>99</td>
<td>0.535</td>
</tr>
<tr>
<td>2003</td>
<td>129</td>
<td>332</td>
<td>203</td>
<td>195</td>
<td>275</td>
<td>80</td>
<td>0.410</td>
</tr>
</tbody>
</table>

Table 4.10: Descriptive Measures of Equity for 6th Grade Reading, 1998-2003
4.3.1.1 Range

The range describes the absolute distance between the minimum and maximum reading scale score each year. Shrinkage in the range statistic implies improvement in horizontal equity. One way of interpreting such shrinkage is that scores are arranged more closely together. Similarly, shrinkage may be interpreted as scores becoming more similar to each other. But such a crude measure is not robust enough to offer this type of information, as we will see next.

The range increased over the period, from 190 in 1998 to a high of 203 in 2003. The expansion in range is explained by a five-point decline in the minimum score (from 134 to 129) and an eight-point increase in the maximum score. The only year in which the range was lower than the 1998 base year was 1999 (range=172). Otherwise, the range generally increased. Changes in the minimum and maximum scores suggest that the observed trend in the range was associated with increases at both ends with the maximum increasing at a greater rate than the minimum. This evidence suggests that equity in reading proficiency scores declined during the study period, i.e., that scores became more separated.

4.3.1.2 Restricted range

In contrast to the range, the restricted range decreased from 1998 ($RR=90$) to 2003 ($RR=80$). The restricted range jumped 13 points between 2001 ($RR=86$) and 2002 ($RR=99$) before falling to the study period low in 2003 ($RR=80$). Contrary to the range statistic, the results for the restricted range suggest that equity in student reading achievement improved rather than declined over the course of the study period.
Examining the upper and lower bounds of the restricted range suggests that the improvement in the restricted range is explained by an eight-point increase in the lower bound, from 187 to 195, and a two-point decline in the upper bound, from 277 to 275, consistent with the principles of egalitarian inequality and the standards-based equity objectives of the state accountability system.

The variation at the boundaries of the restricted range is lower than for the boundaries of the range. Whereas the minimum score varies from 134 in 1998 to 149 in 2002, a difference of 15 points, the lower bound of the restricted range varies by 12 points, from 183 in 1999 to 195 in 2003. The maximum score varies by 27 points, from a low of 319 in 1999 to a high of 346 in 2002 whereas the upper bound of the restricted range varies by only 10 points, from a low of 274 in 1999 to a high of 284 in 2002. The lower variability of the restricted range implies that the range is more susceptible to the influence of outlying scores, a concern raised in the methodological discussion in chapter three, above.

4.3.1.3 Federal range ratio

The reader will recall that the Federal Range Ratio is the quotient of the Restricted Range divided by the score at the fifth percentile. The statistic takes on values from zero to positive infinity. Lower scores imply greater equity.

The Federal Range Ratio mirrors the Restricted Range, a result that is consistent with the fact that the variation in the Restricted Range tends to be most strongly associated with variation at the lower boundary of the Restricted Range, the denominator in the Federal Range Ratio. A distinction between the two measures is that the relative
changes from year to year are somewhat greater for the Federal Range Ratio than for the Restricted Range. For example, the Restricted Range increased from 1998 to 1999 by 1.1% whereas the Federal Range Ratio increased 3.3%. From 1998 to 2003 these figures are -11.1% and -14.8%, respectively. Based on the observed decline in the Federal Range Ratio, we conclude that the statistic suggests overall improvement in equity in reading proficiency in the study district over the study period.

4.3.1.4 Summary of the descriptive measures of horizontal equity

Table 4.11 summarizes the results of the three descriptive measures of horizontal equity, indicating the trend in each statistic and the implied trend in equity. The three measures produced inconsistent results, with the range suggesting a decrease in equity while the restricted range and federal range ratio imply an increase in equity in reading scale scores over the period.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Trend</th>
<th>Implied Trend in Equity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>Increasing</td>
<td>Decreasing</td>
</tr>
<tr>
<td>Restricted Range</td>
<td>Decreasing</td>
<td>Increasing</td>
</tr>
<tr>
<td>Federal Range Ratio</td>
<td>Decreasing</td>
<td>Increasing</td>
</tr>
</tbody>
</table>

Table 4.11: Summary of Descriptive Measures of Horizontal Equity for 6th Grade Reading, 1998-2003
4.3.2 Variance-based measures of horizontal equity

This section presents results for the three variance-based measures of horizontal equity: the standard deviation; the coefficient of variation; and, the standard deviation of logarithms. The section concludes with a brief summary.

<table>
<thead>
<tr>
<th>Year</th>
<th>SD</th>
<th>% Chg from 1998</th>
<th>CV</th>
<th>% Chg from 1998</th>
<th>SDL</th>
<th>% Chg from 1998</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>27.034</td>
<td>-</td>
<td>0.116</td>
<td>-</td>
<td>0.121</td>
<td>-</td>
</tr>
<tr>
<td>1999</td>
<td>27.399</td>
<td>1.4</td>
<td>0.118</td>
<td>1.7</td>
<td>0.122</td>
<td>1.0</td>
</tr>
<tr>
<td>2000</td>
<td>26.419</td>
<td>-2.3</td>
<td>0.115</td>
<td>-1.2</td>
<td>0.115</td>
<td>-5.2</td>
</tr>
<tr>
<td>2001</td>
<td>26.449</td>
<td>-2.2</td>
<td>0.112</td>
<td>-3.6</td>
<td>0.114</td>
<td>-5.6</td>
</tr>
<tr>
<td>2002</td>
<td>30.816</td>
<td>14.0</td>
<td>0.131</td>
<td>12.3</td>
<td>0.132</td>
<td>9.0</td>
</tr>
<tr>
<td>2003</td>
<td>26.221</td>
<td>-3.0</td>
<td>0.110</td>
<td>-5.5</td>
<td>0.113</td>
<td>-6.9</td>
</tr>
</tbody>
</table>

Table 4.12: Dispersion-Based Measures of Equity for 6th Grade Reading, 1998-2003

Table 4.12 provides the results of the three measures, along with changes in each measure from the base year to each subsequent year.

4.3.2.1 Standard deviation

The standard deviation \((SD)\) decreased from 1998, \(SD=27.034\), to 2003, \(SD=26.221\). The evidence that the variance among test scores declined suggests an increase in equity with respect to sixth grade reading proficiency. Figure 4.3 provides a plot of the standard deviations. The figure suggests that, with the exception of the 2002 data \((SD=30.816)\), there does appear to be a general downward trend in standard deviation. The marked increase in the standard deviation in 2002 is explored in
somewhat greater detail in the review of quintile means in the following section on
distributional measures of horizontal equity.

Figure 4.3: Standard Deviation of Reading Scale Scores, 1998-2003

4.3.2.2 Coefficient of Variation

The coefficient of variation \( CV \) is the ratio of the standard deviation to the mean.
Another way of expressing this is that the \( CV \) represents the standard deviation as a
proportion of the mean. Equity increases as \( CV \) approaches zero.

The observed \( CV \) values, presented in Table 4.11, declined from 1998, \( CV=0.116 \),
to 2003, \( CV=0.110 \), suggest that equity among reading scale scores increased for the
period. Figure 4.4 illustrates the overall trend in a plot of the \( CV \) values.
As with the SD results, the CV pattern reveals a marked increase in 2002. Holding aside the results for 2002, the CV plot appears to reveal a stronger negative linear trend than the SD. This finding is consistent with the observed positive trend among the reading scale score means: as the mean increases, ceteris paribus, CV declines. The combination of increasing means and decreasing standard deviations should drive the CV down at a greater rate than would be observed in the SD alone.

4.3.2.3 Standard deviation of logarithms

The standard deviation of logarithms (SDL) decreased from 1998, SDL=0.121, to 2003, SDL=0.113. This evidence suggests an increase in equity with respect to sixth
grade reading proficiency. Figure 4.5 offers a visual depiction of the $SDL$ statistic for each of the six years. The pattern appears to be similar to that of $SD$. The observed decline in $SDL$ of 6.9% from 1998 to 2003 was greater than for either the $SD$ (decline =3.0%) or the $CV$ (decline=5.5%). Compared to the changes in the $SD$ and $CV$, the $SDL$ appears to have been less adversely affected in 1999 and 2002 and to have declined (i.e., improved) more rapidly in 2000, 2001, and 2003. This suggests that although the pattern of the $SDL$ is similar to the $SD$, the $SDL$ appears generally to have been more sensitive to equalizing changes and less sensitive to disequalizing changes in the distribution of the sixth grade reading scale scores.

Figure 4.5: Standard Deviation of Logarithms of Reading Scale Scores, 1998-2003
### 4.3.2.4 Summary of variance-based measures of horizontal equity

Table 4.13 summarizes the results of the three variance-based measures of horizontal equity, indicating the trend in each statistic and the implied trend in equity. The three measures produced generally consistent results, with the standard deviation ($SD$) suggesting a more modest relative change in equity than the coefficient of variation ($CV$) or standard deviation of logarithms ($SDL$). The $CV$ suggested a more generally linear relationship among observed values than $SD$ or $SDL$. All three measures implied a sharp, disequalizing departure in 2002 from the overall equalizing trend.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Trend</th>
<th>Implied Trend in Equity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Deviation</td>
<td>Decreasing</td>
<td>Increasing</td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td>Decreasing</td>
<td>Increasing</td>
</tr>
<tr>
<td>Standard Deviation of Logarithms</td>
<td>Decreasing</td>
<td>Increasing</td>
</tr>
</tbody>
</table>

Table 4.13: Summary of Dispersion-Based Measures of Horizontal Equity for 6th Grade Reading, 1998-2003

### 4.3.3 Distributional Measures of Horizontal Equity

This section presents results for the three distributional measures of horizontal equity: the Gini coefficient, Atkinson index, and Theil statistic. This discussion is prefaced with a quintile analysis and concludes with a brief summary.

#### 4.3.3.1 Quintile analysis

Computing quintile means by year offers a direct, conventional method of disaggregating the distribution of scores (e.g., Lynch, 2003). Applied to student
achievement scores, quintile groupings are established by rank-ordering student scores for a given year and then subdividing the students into five groups of equal size. Thus the lowest quintile includes the lowest-scoring 20% of students and the highest quintile has the highest-scoring 20% of students.

If scores were equal, the means of the five quintiles would be equal and the percentage changes from year to year would be equal. To the extent that quintile means differ, an equalizing trend would yield converging means, as indicated by disproportionately greater increases in the lower quintiles than in the upper quintiles.

Table 4.14 presents the quintile means for reading scores for 1998 through 2003. Figure 4.6 provides a plot of the quintile means. The quintile statistics suggest that average within-quintile gains and losses in this district tended to fluctuate and that, over the period 1998-2003, students in the lowest quintile made the greatest percentage increase (3.89%) while students in the highest quintile made the lowest percentage increase (1.92%). From an equity perspective, given the disparity between the upper and lower quintile, this result is associated with an increase in overall equity.

In our earlier review of overall descriptive statistics for the district, above, we noted an increase in the standard deviation in 2002. The quintile results for 2002 suggest that this increased variation in 2002 was associated with an increase in scores within the top quintile and a simultaneous decrease in scores within the bottom quintile.

The average reading score of the lowest 20% of students in the district dropped from 198.1 in 2001 to 193.6 in 2002. At the same time, the average reading score of the top 20% of students in the district increased from 271.8 in 2001 to 279.5 in 2002. So
although mean performance in the district was practically unchanged between 2001 (235.6) and 2002 (235.7), scores expanded away from the mean in both directions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest</td>
<td>193.7</td>
<td>191.6</td>
<td>194.8</td>
<td>198.1</td>
<td>193.6</td>
<td>201.2</td>
<td>-1.04</td>
<td>0.58</td>
<td>2.28</td>
<td>-0.06</td>
<td>3.89</td>
</tr>
<tr>
<td>Q2</td>
<td>220.3</td>
<td>219.8</td>
<td>217.4</td>
<td>223.4</td>
<td>220.8</td>
<td>225.0</td>
<td>-0.23</td>
<td>-1.34</td>
<td>1.41</td>
<td>0.23</td>
<td>2.14</td>
</tr>
<tr>
<td>Q3</td>
<td>232.5</td>
<td>232.8</td>
<td>228.1</td>
<td>235.7</td>
<td>234.9</td>
<td>239.7</td>
<td>0.11</td>
<td>-1.89</td>
<td>1.37</td>
<td>1.01</td>
<td>3.10</td>
</tr>
<tr>
<td>Q4</td>
<td>245.9</td>
<td>245.5</td>
<td>240.3</td>
<td>248.8</td>
<td>249.9</td>
<td>251.6</td>
<td>-0.17</td>
<td>-2.27</td>
<td>1.17</td>
<td>1.62</td>
<td>2.32</td>
</tr>
<tr>
<td>Highest</td>
<td>268.5</td>
<td>267.7</td>
<td>267.3</td>
<td>271.8</td>
<td>279.5</td>
<td>273.6</td>
<td>-0.30</td>
<td>-0.44</td>
<td>1.25</td>
<td>4.09</td>
<td>1.92</td>
</tr>
</tbody>
</table>

Table 4.14: Quintile Means for Reading Scale Scores, 1998 – 2003

Figure 4.6: Quintile Means for Reading Scale Scores, 1998-2003
There is no evidence to suggest that the results are erroneous, but they are anomalous within the context of the observed six-year trend. The 2002 anomaly might be due to idiosyncratic characteristics of the test form administered in that year (e.g., ordering of item difficulties) or to characteristics within the district (e.g., a correspondingly wider distribution of student ability). The district test administrator was unable to identify any particular local phenomenon that might have been responsible for the anomaly (Peters, 2004). As Table 3.1 reveals, statewide results for the 2002 reading test administration reveal a markedly higher SEM in 2002 (SEM=12.09) than for the other years, averaging 5% higher than the average for the other years. An increase in the SEM signals greater dispersion of scores, suggesting that the increased variability observed in the district data coincide with greater variability across the state. These results imply that the test form administered in 2002 was associated with a somewhat greater degree of variability at the extremes than forms for other years. As previously discussed and as illustrated in Appendix A, greater variability at the extremes is plausible given the increased distances between individual raw scores, a feature that might make these scores somewhat less stable than scores near the cut score. If the increase in variability in 2002 is due to test form characteristics, an alternative approach would be to examine equity results with and without the 2002 data in order to evaluate the results both ways, suspending for the moment any judgment regarding which alternative is ‘right.’ This issue should be explored in greater detail and will be revisited in the recommendations section in Chapter Six.

Representing each quintile’s relative share of each year’s scores offers a standardized (i.e., independent of the score scale) view of the pattern and trend of change
among quintiles (Lynch, 2003). In a perfectly equal distribution of scores, the sum of scores within each quintile will account for 20% of the total sum of scores for a year. For any given initial state of inequality, we are interested to see whether quintile shares have converged. Table 4.15 presents the relative quintile shares for each of the years 1998 through 2003. Figure 4.7 provides a visual illustration of these results.

As the quintile means would suggest, the evidence here shows that the students in the highest quintile earn a disproportionately high share of reading scores (e.g., 23.1% of scores in 1998) while students in the bottom quartile earn a disproportionately low share (e.g., 16.7% of scores in 1998). Over the period, students in the lowest quintile increased their relative share, from 16.7% in 1998 to 16.9% in 2003. The year-to-year changes within each quintile do not suggest a consistent pattern, however. During the intervening years, the relative share in Q1 fell from 16.7 to 16.5 in 1999, rose in 2000 (17.0), lost some ground in 2001 (16.8), fell again in 2002 (16.4), and by 2003 (16.9) recovered most of the share held in 2000. Overall, the change in Q1 implies the possibility of upward movement, which additional future data might more clearly define.

Relative share accounted for by the highest quintile declined, from 23.1% in 1998 to 23.0% in 2003 but again, the change is small and the year-to-year changes offer no evidence of a pattern. Overall, the change in Q5 implies that the share of total scores obtained by the top 20% of students has remained fairly stable. Comparing the results for Q1 and Q5, the weak improvement in Q1 and lack of change in Q5 suggest that, if anything, achievement equity may be improving, but if so only slightly.
<table>
<thead>
<tr>
<th>Student Quintile</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest</td>
<td>16.7</td>
<td>16.5</td>
<td>17.0</td>
<td>16.8</td>
<td>16.4</td>
<td>16.9</td>
</tr>
<tr>
<td>Q2</td>
<td>19.0</td>
<td>19.0</td>
<td>18.9</td>
<td>18.9</td>
<td>18.7</td>
<td>18.9</td>
</tr>
<tr>
<td>Q3</td>
<td>20.1</td>
<td>20.2</td>
<td>19.9</td>
<td>20.1</td>
<td>19.9</td>
<td>20.1</td>
</tr>
<tr>
<td>Q4</td>
<td>21.1</td>
<td>21.2</td>
<td>20.9</td>
<td>21.1</td>
<td>21.2</td>
<td>21.1</td>
</tr>
<tr>
<td>Highest</td>
<td>23.1</td>
<td>23.1</td>
<td>23.3</td>
<td>23.1</td>
<td>23.7</td>
<td>23.0</td>
</tr>
<tr>
<td>Sum</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 4.15: Annual Proportions of Scores by Quintile for 6th Grade Reading, 1998-2003

Figure 4.7: Relative Proportion of 6th Grade Reading Scale Scores By Quintile, 1998-2003
4.3.3.2 Gini coefficient

The Gini coefficient ($G$) represents the ratio of the observed cumulative proportion of ranked scores to the idealized cumulative proportion if all students earned the same score. The statistic varies on the range (0,1), where higher scores represent greater equity. Table 4.15 presents the results for the $G$ for each of the six study years.

Consistent with the descriptive and variance-based measures of equity and with the reported statewide score variability, $G$ declines in 2002. Holding out 2002, the results for $G$ suggest a weak increase, yielding an overall gain of .0031, or 0.33% for the study period. These changes, while small, must be understood in the context of the overall values. Since the maximum value of $G$ is 1.0, we see that the observed values are close to the maximum, limiting the range within which the statistic might increase.

The benchmark value $G'$ represents the equity statistic under the hypothetical scenario in which all students whose scores fell below the reading proficiency cut score are brought up to the cut score. This offers a benchmark representing the minimum level of equity that could be obtained for the observed distribution of scores if all students at least met the proficiency criterion.

The results in Table 4.16 indicate that the gap $G-G'$ decreased from -0.0189 in 1998 to -0.0125 in 2003. The plotted values in Figure 4.8 illustrate that the shrinkage is due to a combination of increasing values of $G$ and decreasing values of $G'$. These results suggest that, while the district has demonstrated at least some progress toward improving equity, six more years under similar conditions would not be adequate to bring the district to the level of equity required by NCLB.
Table 4.16: Gini Coefficient for Sixth Grade Reading Scale Scores, 1998-2003

<table>
<thead>
<tr>
<th>Year</th>
<th>$G$</th>
<th>% Chg from 1998</th>
<th>$G'$</th>
<th>$G-G'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>0.9358</td>
<td>-</td>
<td>0.9547</td>
<td>-0.0189</td>
</tr>
<tr>
<td>1999</td>
<td>0.9346</td>
<td>-0.13</td>
<td>0.9552</td>
<td>-0.0206</td>
</tr>
<tr>
<td>2000</td>
<td>0.9375</td>
<td>0.18</td>
<td>0.9558</td>
<td>-0.0183</td>
</tr>
<tr>
<td>2001</td>
<td>0.9376</td>
<td>0.20</td>
<td>0.9525</td>
<td>-0.0149</td>
</tr>
<tr>
<td>2002</td>
<td>0.9275</td>
<td>-0.89</td>
<td>0.9463</td>
<td>-0.0188</td>
</tr>
<tr>
<td>2003</td>
<td>0.9389</td>
<td>0.33</td>
<td>0.9514</td>
<td>-0.0125</td>
</tr>
</tbody>
</table>

Figure 4.8: Observed and Benchmark Gini Coefficient for 6th Grade Reading Scale Scores, 1998-2003
With the exception of the 2002 results, the overall trend of $G$ for the period appears to be weakly positive. The observed values of $G$ appear, again with the exception of 2002, to converge with the benchmark value of $G'$. The results possibly suggest a weak trend of improvement. The weakness of the trend, however, must be emphasized due both to the unevenness in year-to-year shifts, and to the fact that the 1998 data include a smaller $n$ and the 2002 data are associated with a statewide increase in variability.

### 4.3.3.3 Atkinson index

The Atkinson index, $A$, provides a measure of equity that directly incorporates a social welfare dimension, suggesting an interpretation of results in terms of change in social utility or well-being. The Atkinson index is distinguished for its emphasis on changes on the left side of the distribution (lower scoring students), yielding greater improvement for a unit of improvement toward the mean from the left than for an equivalent unit shift toward the mean from the right. From the perspective of performance equity, this translates into an incentive to achieve equity by raising the performance of students on the low end rather than by a mixture of improvement on the left and reduced effort on the right.

Another distinctive feature of $A$ is the coefficient $e$, representing inequity aversion on the left end of the distribution. Higher values of $e$ would reflect greater sensitivity to inequity in the lower end, particular among extreme scores. Following the current literature in the study of income inequality, this study adopts the conventional level of $e=.5$ (e.g., Lynch, 2003) as the aversion factor.
Results for $A$, reported in Table 4.17 and illustrated in Figure 4.9, suggest a similar pattern to $G$. The values of $A$ for the years 1999 (.9964), 2000 (.9967), 2001 (.9968), and 2003 (.9969) suggest a weakly positive trend. As with $G$, the observed values of $A$ appear, with the exception of 2002, to converge with the benchmark value of $A^\prime$. The results possibly suggest a weak trend of improvement. As noted above in the discussion of the Gini coefficient, the weakness of the trend must be emphasized due both to the unevenness in year-to-year shifts and also to the concerns with the 1998 and 2002 data.

<table>
<thead>
<tr>
<th>Year</th>
<th>$A_{0.5}$</th>
<th>% Chg from 1998</th>
<th>$A'_{0.5}$</th>
<th>$A_{0.5}-A'_{0.5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>0.9965</td>
<td></td>
<td>0.9983</td>
<td>-0.0018</td>
</tr>
<tr>
<td>1999</td>
<td>0.9964</td>
<td>-0.01</td>
<td>0.9983</td>
<td>-0.0019</td>
</tr>
<tr>
<td>2000</td>
<td>0.9967</td>
<td>0.02</td>
<td>0.9982</td>
<td>-0.0015</td>
</tr>
<tr>
<td>2001</td>
<td>0.9968</td>
<td>0.03</td>
<td>0.9982</td>
<td>-0.0014</td>
</tr>
<tr>
<td>2002</td>
<td>0.9957</td>
<td>-0.08</td>
<td>0.9976</td>
<td>-0.0019</td>
</tr>
<tr>
<td>2003</td>
<td>0.9969</td>
<td>0.04</td>
<td>0.9981</td>
<td>-0.0012</td>
</tr>
</tbody>
</table>

Table 4.17: Atkinson Index for 6th Grade Reading Scale Scores, 1998-2003
Unlike the equity literature that has focused on income distributions, the equity literature in educational finance suggests that an inequality-aversion coefficient value of \(e=0.5\) represents a low sensitivity to relative differences on the left side of the distribution (Berne & Steiffel, 1984). Values of \(e=30\) or more are described in that literature as large (ibid). Comparing results with a larger value of \(e\) reveals the sensitivity of the statistic to the inequity aversion component.

Results for \(A_{30}\), reported in Table 4.18 and illustrated in Figure 4.10, reveal a distinctly different pattern in equity for the study period. While \(A_{0.5}\) and \(G\) exhibit a slight net improvement in equity, \(A_{30}\) suggests a decrease in equity over the study period.
with increases rather than decreases in 1999 and 2002. Given the greater sensitivity of \( e \) to changes in the low end of the distribution of student scores, these results suggest that the increased inequity detected at \( A_{0.5} \) in 1999 and 2002 were characterized by greater inequity at the upper end of the distribution and less inequity at the lower end of the distribution.

Exploring this issue further, the quintile analysis might be expected to reveal corroboration for this latter inference. That is, if the lowest quintile lost ground against the highest quintile, then the results of \( A_{30} \) would be supported. Looking at the quintile results in Table 4.13, however, the results appear to be just the opposite: the lower quintiles appear to have gained ground on the upper quintiles, consistent with the indications of \( A_{0.5} \).

The literature offers no guidance on sensitivity analysis for outliers on \( A \). One possible explanation of the observed results is that raising the coefficient \( e \) to the level of 30 makes \( A_{30} \) overly sensitive to the positions of outlying scores. Returning to the individual student scores, the minimum score in 1998 was 134 while in 2003 the minimum was 129. One simple approach to evaluating the sensitivity of \( A_{30} \) to outlying reading scores would be to reduce the lowest score by some number of points and check the impact on \( A_{30} \).
Using the observed scores as a frame of reference suggests adjusting the minimum score in 2003 (129) to the level observed in 1998 (134) and rechecking $A_{30}$.

The result of this shift drives $A_{30}$ from the observed result of $A_{30}=0.6840$ to $A_{30}=0.7102$. 
This change in a single reading score, from 129 to 134, accounts for 82% of the
difference between the computed results of $A_{30}$ in 1998 and 2003. These findings call
into question the values for $A_{30}$, particularly in the face of the observed disequalizing
characteristics of the 2002. Raising the coefficient to $e=30$ appears to have so greatly
increased sensitivity to changes in the lower end of the distribution as to destabilize $A$.
Such volatility therefore renders interpretation of the results for $A_{30}$ uncertain.

4.3.3.4 Theil statistic

The Theil statistic, $T$, provides a measurement of equity in terms of the physical
principle of entropy such that equality among scores would represent a condition of
uniformity and any differences among scores would represent a departure from such
uniformity (Theil, 1967). The statistic has a minimum of zero, representing perfect
equity. There is no positive maximum value for $T$.

The results for $T$, presented in Table 4.19 and plotted in Figure 4.11, suggest
results analogous to the findings for $G$ and $A$, above. Since zero is the ‘optimum’ value
from an equity perspective, a negative trend in $T$ suggests improvement in equity. The
Theil statistic reveals the same marked decrease in equity in 2002 observed in the other
distributional equity measures; in this case the value for $T$ increased by 23.62% in 2002
over the base year, 1998, in contrast to improvements of 4.96% in 2000, 8.20% in 2001,
and 11.40% in 2003. If we hold out the results for 2002, the changes might suggest a
weak trend of improved equity, though the unevenness from year to year suggests that
such an interpretation be taken with caution.
<table>
<thead>
<tr>
<th>Year</th>
<th>T</th>
<th>% Chg from 1998</th>
<th>T’</th>
<th>T-T’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>0.0069</td>
<td>-</td>
<td>0.0034</td>
<td>0.0035</td>
</tr>
<tr>
<td>1999</td>
<td>0.0072</td>
<td>4.19</td>
<td>0.0034</td>
<td>0.0038</td>
</tr>
<tr>
<td>2000</td>
<td>0.0066</td>
<td>-4.96</td>
<td>0.0037</td>
<td>0.0029</td>
</tr>
<tr>
<td>2001</td>
<td>0.0063</td>
<td>-8.20</td>
<td>0.0037</td>
<td>0.0026</td>
</tr>
<tr>
<td>2002</td>
<td>0.0085</td>
<td>23.62</td>
<td>0.0049</td>
<td>0.0036</td>
</tr>
<tr>
<td>2003</td>
<td>0.0061</td>
<td>-11.40</td>
<td>0.0038</td>
<td>0.0023</td>
</tr>
</tbody>
</table>

Table 4.19: Theil Statistic for 6th Grade Reading Scale Scores, 1998-2003

With respect to $T'$, the simulated benchmark equity target for $T$, the distributions of reading scale scores appear to have made progressive improvement, closing the gap.
between $T$ and $T'$ from 0.0035 in 1998 to 0.0023 in 2003. A visual representation of this relationship is offered in the converging values of $T$ and $T'$ in Figure 4.11.

4.3.3.5 Summary of the distributional measures of horizontal equity

The results for the three measures, $G$, $A$, and $T$, follow a generally similar pattern as is evident in the plots in Figures 4.8, 4.9, and 4.11, respectively. A prominent point of agreement among these graphs is the convergence with the benchmark statistics ($G'$, $A'$, and $T'$, respectively), a relatively large negative shift in the statistics in the year 2002, and what may be a described as very weak improvement in equity across the remaining study years.

Table 4.20 summarizes the results of the three distributional measures of horizontal equity, indicating the trend in each statistic and the implied trend in equity. The quintile analysis provided a descriptive tool for preliminary examination of the score distributions. The results helped illustrate the disequalizing shift in 2002, suggested by the increased score variability reported for the statewide test administration and indicated across $SD$, $CV$, $SDL$, $G$, $A$, and $T$. The quintile results revealed that the shift is associated with a general expansion of the entire score distribution. Reviewing the data provided by the Department of Education, Table 3.1, above, this shift may have been due to the functional characteristics of the particular version of the test instrument administered in 2002. Given this possibly confounding effect, results of the equity analyses were described both with and without the 2002 data.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Trend</th>
<th>Implied Trend in Equity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quintile Analysis (Q1 vs Q5)</td>
<td>Uneven; Marked divergence in 2002.</td>
<td>Unclear overall. When holding out 2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>results, pattern suggests weak improvement.</td>
</tr>
<tr>
<td>Gini Coefficient</td>
<td>Uneven; Convergence with benchmark, $G'$</td>
<td>Unclear overall. When holding out 2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>results, pattern suggests weak improvement.</td>
</tr>
<tr>
<td>Atkinson Index for $e=0.5$</td>
<td>Uneven; Convergence with benchmark, $A'$</td>
<td>Unclear overall. When holding out 2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>results, pattern suggests weak improvement.</td>
</tr>
<tr>
<td>Atkinson Index for $e=30$</td>
<td>Uneven; Contradicts $G$, $A$, $T$;</td>
<td>Too volatile to interpret</td>
</tr>
<tr>
<td></td>
<td>宋敏strong to single outliers</td>
<td></td>
</tr>
<tr>
<td>Theil Statistic</td>
<td>Uneven; Convergence with benchmark, $A'$</td>
<td>Unclear overall. When holding out 2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>results, pattern suggests weak improvement.</td>
</tr>
</tbody>
</table>

Table 4.20: Summary of Distributional Measures of Horizontal Equity for 6th Grade Reading Scale Scores, 1998-2003

The results for $G$, $A$ at $e=0.5$, and $T$ were quite comparable, revealing small net improvements for the entire study period but characterized by uneven annual changes, particularly for 2002. Holding out 2002, the results for all three measures appear to suggest the possibility of a weak trend of improvement. The limited number of data points, the small absolute sizes of the changes, and the annual volatility caution against the over-interpretation of these findings. Examining the stability of results for $A$ at different values of $e$ suggests that, while increases in $e$ make $A$ more sensitive to changes at the low end of the distribution, the statistic can be strongly influenced by even a single outlying score, rendering the measure too volatile for reliable use.
4.4 Research Question 3: Have reading scores converged within and between schools?

This section examines the results of the analysis of within- and between-schools contributions to school district equity in reading scale scores between 1998 and 2003. The section presents results of the factorial decomposition of the Atkinson index and the additive decomposition of the Theil statistic. Evidence is evaluated with respect to indications of whether relative shares of within school equity and between school equity have remained stable over the period, suggesting, in light of the equalizing linear trend at the district level, convergence both within and between schools.

4.4.1 Descriptive statistics by school

Table 4.21 presents descriptive statistics for reading scale scores by school for the years 1998-2003. We note first the variability in student totals for schools across years. Some schools reveal large percentage gains and losses from year to year (e.g., School 8). The district administration (Peters, 2004) suggested that the differences were likely due to factors including normal changes in student enrollment, annual volatility in busing and neighborhood attendance zones decided each year in order to balance students across schools, and incomplete student records due to transfers of student files between elementary and middle schools. The administrator did not believe that these variations would be associated with systematic factors such as race, income, sex, or other student characteristic that might affect the interpretability of results.

Means are plotted in Figures 4.12 through 4.23. Most schools, with the exception of Schools 2 and 6, demonstrated net gains in mean score during the period. In terms of
between-schools variability, the means appear to be separated fairly widely, e.g., School 4 vs School 6 in 1998 (mean difference = 18.7 scaled score points). The plots of mean reading scale scores, presented against a common vertical scale, help illustrate this feature. The standard deviations of reading scores within schools are large in comparison with mean differences between schools. For example, School 4 standard deviation in 1998 is 28.5 and in School 6, 27.6, suggesting greater variability within schools than between schools. Between school differences in standard deviation are noted within each year. For example, in 1998 the standard deviation ranges from a low of 23.717 for School 1 to a high of 31.655 for School 12. Relative rankings on standard deviation among the schools change substantially from year to year, with most schools being among the highest and lowest standard deviations in different years. Departing from this generally random pattern, Schools 10 and 11 tend consistently to have low standard deviations relative to the other schools. School 12, meanwhile, tends consistently to have a higher standard deviation relative to the other schools.

We noted above that in the year 2002, the district as a whole demonstrated greater variation among scores, consistent with the observed increase in variance across the state that year. Results for the individual schools for the year 2002 suggest that the increased variance seen at the district level extends across all of the schools. Looking at School 1 in Table 4.25, for example, the standard deviation increases from 25.005 in 2001 to 31.426 in 2002, falling back to 23.978 in 2003.

The decomposition of equity statistics provides a method of measuring within- and between-schools equity each year. In order to gain a preliminary understanding of the equity decomposition, a useful quantification of within- and between-schools
variation are the variance components of a one-way ANOVA of reading scale scores by schools for each of the study years. Analogous to the equity decomposition for each year, the ANOVA approach yields the mean squares associated with between-schools variation each year and the mean squares associated with within-schools variation for each year. This information will assist in understanding the basic structure of the variance each year and will offer a preview of the equity analysis. Table 4.22 summarizes the results of the analysis. The focus here is not on hypothesis testing or the $F$ statistic, but rather on the descriptive information offered via the mean-squares ($MS$).

Between-schools variance, $MS_{between}$, declined in 1999 and then gained at unequal intervals in 2000, 2001, and 2003, with the exception again of 2002. Within-schools variance, $MS_{within}$, follows a pattern opposite to that of $MS_{between}$, rising in 1999 and declining unevenly thereafter in 2000, 2001, and 2003, again excepting the results of 2002. While we are not directly interested in the results of the individual $F$ tests, the results suggest that there is a statistically significant amount of between-schools variation each year. The ratio of between-schools to within-schools variation, represented by the $F$ statistic, suggests that while the between-schools portion yielded a net increase from 1998 to 2003, the annual variability suggests the absence of any clear pattern. Overall, the mean-squares data suggest that while there was a net decrease in within-schools variance and a net increase in between-schools variance, the changes have been very uneven. The net changes suggest that equity improved within schools and declined between schools. Again, it is important to consider the pattern holding 2002 out of the analysis. Graphs of mean reading scores, Figures 4.12 – 4.23, help illustrate both the annual variability and the school-to-school variability over the period.
<table>
<thead>
<tr>
<th>Sch.</th>
<th>1998</th>
<th></th>
<th>1999</th>
<th></th>
<th>2000</th>
<th></th>
<th>2001</th>
<th></th>
<th>2002</th>
<th></th>
<th>2003</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>36</td>
<td>239.7</td>
<td>23.717</td>
<td>67</td>
<td>237.5</td>
<td>28.480</td>
<td>77</td>
<td>229.8</td>
<td>23.421</td>
<td>60</td>
<td>243.5</td>
<td>25.005</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>234.7</td>
<td>26.826</td>
<td>57</td>
<td>229.8</td>
<td>24.014</td>
<td>55</td>
<td>228.6</td>
<td>22.416</td>
<td>53</td>
<td>233.7</td>
<td>27.442</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>229.5</td>
<td>24.286</td>
<td>68</td>
<td>231.8</td>
<td>25.167</td>
<td>79</td>
<td>229.5</td>
<td>22.829</td>
<td>74</td>
<td>237.0</td>
<td>28.818</td>
</tr>
<tr>
<td>4</td>
<td>77</td>
<td>220.6</td>
<td>28.541</td>
<td>94</td>
<td>223.0</td>
<td>28.321</td>
<td>82</td>
<td>220.6</td>
<td>24.562</td>
<td>76</td>
<td>226.0</td>
<td>26.459</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>229.3</td>
<td>27.554</td>
<td>67</td>
<td>228.9</td>
<td>21.280</td>
<td>78</td>
<td>222.2</td>
<td>28.779</td>
<td>42</td>
<td>233.0</td>
<td>25.475</td>
</tr>
<tr>
<td>6</td>
<td>75</td>
<td>239.3</td>
<td>27.624</td>
<td>80</td>
<td>233.9</td>
<td>26.960</td>
<td>91</td>
<td>238.1</td>
<td>30.384</td>
<td>67</td>
<td>238.7</td>
<td>22.910</td>
</tr>
<tr>
<td>7</td>
<td>79</td>
<td>230.4</td>
<td>27.356</td>
<td>85</td>
<td>233.2</td>
<td>31.549</td>
<td>81</td>
<td>234.9</td>
<td>24.328</td>
<td>78</td>
<td>237.9</td>
<td>29.037</td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>222.2</td>
<td>24.667</td>
<td>45</td>
<td>227.9</td>
<td>27.770</td>
<td>48</td>
<td>223.4</td>
<td>27.961</td>
<td>42</td>
<td>226.8</td>
<td>25.529</td>
</tr>
<tr>
<td>9</td>
<td>91</td>
<td>237.5</td>
<td>27.597</td>
<td>85</td>
<td>237.7</td>
<td>31.066</td>
<td>89</td>
<td>232.0</td>
<td>26.899</td>
<td>83</td>
<td>241.4</td>
<td>26.892</td>
</tr>
<tr>
<td>10</td>
<td>49</td>
<td>236.9</td>
<td>20.634</td>
<td>55</td>
<td>231.0</td>
<td>25.406</td>
<td>69</td>
<td>231.7</td>
<td>25.702</td>
<td>55</td>
<td>243.3</td>
<td>21.394</td>
</tr>
<tr>
<td>11</td>
<td>37</td>
<td>237.4</td>
<td>25.285</td>
<td>39</td>
<td>236.6</td>
<td>22.210</td>
<td>50</td>
<td>229.0</td>
<td>25.261</td>
<td>58</td>
<td>231.0</td>
<td>24.420</td>
</tr>
<tr>
<td>12</td>
<td>27</td>
<td>228.0</td>
<td>31.655</td>
<td>44</td>
<td>225.8</td>
<td>25.849</td>
<td>41</td>
<td>231.9</td>
<td>29.010</td>
<td>68</td>
<td>230.9</td>
<td>25.545</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>676</td>
<td>232.2</td>
<td>27.034</td>
<td>786</td>
<td>231.5</td>
<td>27.399</td>
<td>840</td>
<td>229.6</td>
<td>26.419</td>
<td>756</td>
<td>235.6</td>
</tr>
</tbody>
</table>

Table 4.21: Descriptive Statistics for Reading Scale Scores by School, 1998-2003
Figure 4.12: Mean Reading Scale Score for School 1, 1998-2003

Figure 4.13: Mean Reading Scale Score for School 2, 1998-2003
Figure 4.14: Mean Reading Scale Score School 3, 1998-2003

Figure 4.15: Mean Reading Scale Score School 4, 1998-2003
Figure 4.16: Mean Reading Scale Score for School 5, 1998-2003

Figure 4.17: Mean Reading Scale Score for School 6, 1998-2003
Figure 4.18: Mean Reading Scale Score for School 7, 1998-2003

Figure 4.19: Mean Reading Scale Score for School 8, 1998-2003
Figure 4.20: Mean Reading Scale Score for School 9, 1998-2003

Figure 4.21: Mean Reading Scale Score for School 10, 1998-2003
Figure 4.22: Mean Reading Scale Score for School 11, 1998-2003

Figure 4.23: Mean Reading Scale Score for School 12, 1998-2003
<table>
<thead>
<tr>
<th>Year</th>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>Between</td>
<td>26093</td>
<td>11</td>
<td>2372</td>
<td>3.371*</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>467221</td>
<td>664</td>
<td>704</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>493313</td>
<td>675</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>Between</td>
<td>16855</td>
<td>11</td>
<td>1532</td>
<td>2.072*</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>572464</td>
<td>774</td>
<td>740</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>589320</td>
<td>785</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Between</td>
<td>22761</td>
<td>11</td>
<td>2069</td>
<td>3.044*</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>562824</td>
<td>828</td>
<td>680</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>585584</td>
<td>839</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Between</td>
<td>24378</td>
<td>11</td>
<td>2216</td>
<td>3.273*</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>503791</td>
<td>744</td>
<td>677</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>528169</td>
<td>755</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Between</td>
<td>28959</td>
<td>11</td>
<td>2633</td>
<td>2.839*</td>
<td>.036</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>767798</td>
<td>828</td>
<td>927</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>796758</td>
<td>839</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Between</td>
<td>26210</td>
<td>11</td>
<td>2383</td>
<td>3.577*</td>
<td>.043</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>581572</td>
<td>873</td>
<td>666</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>607782</td>
<td>884</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: * significant at .025 level.

Table 4.22: Summary of One-Way ANOVA of Reading Scale Scores by Schools, 1998-2003

### 4.4.2 Atkinson Decomposition

Following the technique detailed by Lasso de la Vega & Urrutia (2003), the factorial decomposition of the Atkinson index permits examination of the within- and between-schools sources of equity from the overall district distribution of reading scale scores. As a factorial decomposition of fractional values on the range (0,1), a larger value here represents a larger contribution to equity. Table 4.23 presents the results of the decomposition. Figure 4.24 offers a graphical representation of the values. The
reader will note that the scales on the figure differ by a factor of 10, a condition which enables the plotting of both sets of values while maintaining a narrow enough range on each scale to discern the relationships among individual points.

Consistent with the findings in the review of descriptive statistics, the results of the Atkinson index at $A_{0.5}$ suggest that between-schools equity is greater than within-schools equity. For example, the 1998 value for $A_{0.5\text{between}}$ (0.9998) is greater than that of $A_{0.5\text{within}}$ (0.9967). This ordering holds across all years, 1998-2003.

The relationship between the within-schools component and the overall statistic suggests that the within-schools component dominates the overall equity, a suggestion confirmed by the relative sizes of the factorial components each year. The trend in equity components generally conforms to the review of descriptive statistics and mean squares. The between-schools component rises, i.e., implying increased equity, in 1999 and subsequently declines in 2000, 2001, and 2002 before rising again in 2003. The within-schools component follows a pattern consonant to the statistic for the district as a whole, declining in 1999, and rising in 2000, 2001, and 2003, with an intervening decline in 2002. Overall, no strong pattern is evident for either component. Changes in between-schools equity appear to cancel on a bumpy but flat line. Within-schools equity, too, appears inconsistent from year to year. If we withhold 2002 from the within-schools analysis, the pattern appears to conform to the unstable but weakly positive pattern evident in the district-level analyses for $G$, $A$, and $T$, above.
Table 4.23: Atkinson Decomposition of Within- and Between-Schools Equity, 1998-2003

<table>
<thead>
<tr>
<th>Year</th>
<th>$A_{0.5}^{between}$</th>
<th>$A_{0.5}^{within}$</th>
<th>$A_{0.5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>0.99982</td>
<td>0.99668</td>
<td>0.99650</td>
</tr>
<tr>
<td>1999</td>
<td>0.99990</td>
<td>0.99651</td>
<td>0.99641</td>
</tr>
<tr>
<td>2000</td>
<td>0.99987</td>
<td>0.99685</td>
<td>0.99672</td>
</tr>
<tr>
<td>2001</td>
<td>0.99985</td>
<td>0.99696</td>
<td>0.99681</td>
</tr>
<tr>
<td>2002</td>
<td>0.99984</td>
<td>0.99587</td>
<td>0.99572</td>
</tr>
<tr>
<td>2003</td>
<td>0.99987</td>
<td>0.99705</td>
<td>0.99691</td>
</tr>
</tbody>
</table>

Figure 4.24: Atkinson Decomposition of Within- and Between-Schools Equity of 6th Grade Reading Scale Scores, 1998-2003
4.4.3 Theil Decomposition

The additive decomposition of the Theil statistic, following Conçeição & Ferreira (2000), permits isolation of within- and between-schools equity components such that the two values sum to the overall Theil statistic observed for a given distribution of reading scale scores. Table 4.24 presents the results for the Theil decomposition of reading scale scores for the study district for the years 1998-2003. The additive nature of the decomposition permits evaluation of the components in proportion to one another and the total. Table 4.28 provides the percent of total inequity associated with the within-schools component. The percent of inequity associated with the between-schools component is the difference between the within schools percentage and 100. Figure 4.25 provides a visual representation of the relationship among the decomposed values. The reader will note that the scales on the left and right, corresponding to the within- and between-schools components, respectively, differ by a factor of 10. The reader is reminded that the minimum value of the Theil statistic is zero, corresponding to perfect equity. Values greater than zero representing increasing levels of inequity.

The results of the Theil decomposition for reading scale score distributions are consistent with the Atkinson decomposition. Between-schools equity improves in 1999 and declines in 2000, 2001, and 2002 before improving again in 2003. The between-schools component of the Theil statistic accounts for 5.2% of the value of $T$ in 1998 and 4.3% of $T$ in 2003, suggesting that between-schools equity is better by a factor of 10 when compared to within-schools equity. The within-schools portion of equity, accounting for 94.8% of the observed value of $T$ in 1998, dominates the districtwide $T$ values.
### Table 4.24: Theil Decomposition of Within- and Between-Schools Equity of 6th grade reading scale scores, 1998-2003

<table>
<thead>
<tr>
<th>Year</th>
<th>$T$</th>
<th>$T_{between}$</th>
<th>$T_{within}$</th>
<th>% Within</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>0.0069</td>
<td>0.00036</td>
<td>0.0065</td>
<td>94.8</td>
</tr>
<tr>
<td>1999</td>
<td>0.0072</td>
<td>0.00020</td>
<td>0.0070</td>
<td>97.2</td>
</tr>
<tr>
<td>2000</td>
<td>0.0066</td>
<td>0.00026</td>
<td>0.0063</td>
<td>96.1</td>
</tr>
<tr>
<td>2001</td>
<td>0.0063</td>
<td>0.00029</td>
<td>0.0061</td>
<td>95.4</td>
</tr>
<tr>
<td>2002</td>
<td>0.0085</td>
<td>0.00031</td>
<td>0.0082</td>
<td>96.3</td>
</tr>
<tr>
<td>2003</td>
<td>0.0061</td>
<td>0.00026</td>
<td>0.0059</td>
<td>95.7</td>
</tr>
</tbody>
</table>

Overall, no strong pattern is evident for either component. Changes in between-schools equity appear to cancel, as in the Atkinson decomposition, along a bumpy but
otherwise flat line. Within-schools equity, too, appears unstable from year to year. If we withhold 2002 from the within-schools analysis, the pattern appears to conform to the weakly positive pattern evident in the Atkinson decomposition and in the district-level analyses for \( G \), \( A \), and \( T \), above.

### 4.4.4 Summary of the within- and between-schools equity decompositions

The decomposition of the Atkinson and Theil statistics permitted separate evaluation of the contributions to equity of between-schools and within-schools distributions of reading scale scores for the study district for the period 1998-2003. The basic descriptive data, i.e., school means and standard deviations, together with the variance components of the analysis of variance offered evidence that equity is greater between the schools than within the schools. The factorial decomposition of the Atkinson index and the additive decomposition of the Theil index were consistent with the indications of the descriptive statistics. Of the two equity components, i.e., within-schools and between-schools, the within-schools component reveals much less equity, a logical result considering the range of scores within a school compared with the much smaller overall differences between schools. Examining each component separately, the findings suggest that the within-schools pattern reflects the district-level results, revealing a small net improvement in equity if the volatile 2002 data are held out. At the between-schools level, the equity statistic does not offer a discernable pattern. In sum, then, the results suggest that scores moved closer together within schools, but that schools did not move closer together relative to each other.
CHAPTER 5
REVIEW OF FINDINGS

5.1 Introduction

This study has provided an analysis of scores from the Ohio sixth grade proficiency test of reading for students of a large, Midwestern school district for each of the years 1998 through 2003. The analysis informed three research questions relating to the characteristics of student performance over the period. The work focused on change in the equality of student achievement over the period in the district overall and within and between schools. This chapter provides a summary of the analytical findings. The following chapter offers a discussion of suggestions implied in those findings, the limitations of the analysis and interpretations, and suggestions for future research.

5.2 Summary of findings

This section opens with a review of violations of assumptions and proceeds to a summary of findings with respect to each of the three research questions and associated statistical hypothesis tests.
5.2.1 Research Question 1: Has reading proficiency increased in the district?

The study examined the trend in sixth-grade reading proficiency in the study district over the six-year period, 1998 through 2003. The first method of evaluation focused on the generally accepted practice of determining change based on reading proficiency rates. The second method of evaluation offered an analysis of variance of mean reading proficiency scores over the period. Follow-up tests included the Dunnet test, comparing each year’s mean to the base year, 1998, the Scheffé test, comparing each year’s mean to that of all other year’s, in order to explore the appropriateness of 1998 as the base year, and the test of linear trend in order to determine whether the pattern of means generally conformed to a linear improvement trend.

5.2.1.1 Pass rates and descriptive statistics

Approaching the problem from the perspective of current practice in state accountability systems, the percent of students scoring at or above the state cut score for reading proficiency suggested an overall net gain for the study period but no clear pattern of year-to-year performance, with proficiency rates declining from 1998 to 2000, rising in 2001, and declining again somewhat in 2002 before peaking in 2003. The results suggested that the low degree of general change in scores may have explained the annual volatility of the pass rates.

A concern raised in the review of descriptive statistics was the variation in numbers of students reported by the district each year. While the district administrator responsible for testing noted that most of the annual variation was within a range consistent with grade-to-grade and year-to-year fluctuations, the number of records
available for 1998 represented a substantial volume of missing records due, apparently, to certain early inconsistencies in the electronic file maintenance processes. Based on the district’s statistics, the administrator reported that the data for 1998 appeared to be consistent with other years in terms of student characteristics, supporting the suggestion that the missing records were missing at random. Rather than using 1998 as a base year, as we had intended for example in the ANOVA post hoc comparisons, the analysis was expanded sufficiently to explore comparisons across all years. Throughout the analyses, the 1998 data appeared to fit reasonably well within the context of the other years’ data.

The second approach to the research question focused on the observed distribution of reading scale scores each year. Looking first at descriptive statistics, the pattern of reading scale score means revealed declines in 1999 and 2000 but suggested a general pattern of improvement from 2000 through 2003. Consistent with the research hypothesis, the descriptive statistics implied a positive trend in mean score. The one-way analysis of variance, ANOVA, permitted examination of the scale score data based on statistical tests of significance of equality of means and linear trend.

5.2.1.2 Assumption checks for ANOVA

The check of assumptions underlying ANOVA revealed certain caveats. Reading scale scores are interval level data, fulfilling the level of measure assumption. Students were not randomly selected or assigned across schools in the study district, violating the assumption of randomization. The violation of randomization limits the generalizability of the study and instructs replication to determine how the measures apply in other contexts (other districts, schools, times). The assumption of independence is indirectly
assured through test security, the fact that the reading test represents cumulative rather than grade-specific knowledge and skills, the fact that students are in the third quarter of sixth grade, suggesting that teacher effects might be muted, and the fact that the district has emphasized instruction based on the state curriculum models since the models were developed in the mid 1990’s, thereby standardizing to a certain degree the domain coverage measured on the test. Given the limited evidence regarding the independence violation, results must be interpreted with caution while reiterating the need for replication studies to address the question of consistency of results in similar and different settings.

Evidence drawn from the Shapiro-Wilk test of normality and, separately, from tests of skewness implied violations of the assumption of normality of underlying population distributions. Although the $F$ distribution tends to be robust with respect to non-normality of symmetric distributions, violations due to asymmetry, i.e., skewness, tend to bias $F$ in the positive direction (Keppel, 1991). Following Keppel (pp.97-98), we reduced the nominal level of confidence from alpha=.05 to alpha=.025.

Evidence from the $F$-max test of homoscedasticity implied that the distribution of observed reading scale scores did not violate the fifth assumption of the one-way ANOVA, equality of underlying population variances.

5.2.1.3 Results of ANOVA

The overall results of the omnibus $F$ test of differences between means for six years, 1998 through 2003, revealed that at least one of the means was significantly different from the others. Three sets of follow up tests explored the nature of these
differences. Dunnett pairwise post hoc comparisons among reading scale score means for each year compared to the base year, 1998, revealed that the observed increases in mean scores were associated with a statistically significant difference by the year 2003. The Scheffé test for all possible single comparisons revealed that the 1998 results had provided a reasonable base year plus additional information regarding significant differences between years. These results revealed that the peak reached in 2003 represented a statistically significant gain over 1998, 1999, and 2000. Otherwise, the information reinforced the indication that the differences between years tended to be negligible and followed no consistent, significant pattern. Results of the post hoc test of linear trend implied that the pattern of means conformed to a statistically significant but substantively trivial positive linear trend. Logic and lack of adequate remaining effect variance deterred continuation to higher order trends.

5.2.1.4 Summary of results for Research Question 1

The analysis of evidence regarding Research Question 1 yielded mixed results regarding the research hypothesis that achievement in the district has improved over the study period. The reading pass rate increased over the period when comparing 1998 with 2003, but no definitive trend was evident across the intervening years. A more detailed examination of the distributions of reading scores for the six year period, by means of one-way ANOVA, suggested that changes in annual reading scale scores tended to follow a weakly significant, positive linear trend, lacking sufficient effect size to be of interest.
5.2.2 Research Question 2: Have reading scores converged across the district?

The study examined the second research question based on evidence derived from three sets of measures of horizontal equity. The measures offered evidence regarding changes at given points in the score distributions, evidence regarding general convergence across the distributions, and evidence regarding convergence weighted toward increases on the left side of the distribution, i.e., among low scoring students.

5.2.2.1 Descriptive statistics

Consistent with the research hypothesis, the descriptive statistics suggested a negative trend in variance among reading proficiency scale scores. An examination of the standard deviation over the study period suggested declining score variance, with an exception noted in 2002. Evidence from the Ohio Department of Education, i.e., annual SEM’s for the reading tests for the six years in the study period, suggested that the increased variance observed in the district in 2002 was associated with a statewide phenomenon that might have been related to characteristics of the particular test form administered in that year. Given this idiosyncratic difference, combined with the ODE data, the results of the subsequent analyses were carefully examined based on the lingering possibility that the 2002 data may have been structurally inconsistent with the data for the other five years.

5.2.2.2 Point measures of horizontal equity

The three point measures of horizontal equity used in the study, the range, restricted range, and federal range ratio, yielded mixed results. Changes in the range
statistic over the period implied a decrease in horizontal equity while the restricted range and federal range ratio implied an increase in horizontal equity. Consistent with concerns raised in the literature (e.g., Sen, 1997), volatility among outlying scores in the distributions, driving the range up in this case, appeared to explain the disagreement among the measures.

5.2.2.3 Variance-based measures of horizontal equity

Results for the three variance-based measures of horizontal equity, the standard deviation ($SD$), coefficient of variation ($CV$), and standard deviation of logarithms ($SDL$), conformed to the research hypothesis that horizontal equity in reading proficiency scale scores increased over the study period. Equity declined sharply for all three measures in 2002, a phenomenon that may have been associated with generally increased statewide variability in the reading test form administered in that year (Ohio Department of Education, 2004; Peters, 2004). Results for the $SDL$ were proportionately larger than for the $CV$ or $SD$. The result implied that at least some of the increase in horizontal equity was due to improvements among students at the low end of the achievement distribution.

5.2.2.4 Distributional measures of horizontal equity

Before examining the distributional measures directly, a quintile analysis offered a preliminary look at distribution-wide changes in reading scale scores over the study period. Results for the top and bottom quintiles suggested that reading achievement at the low end of the distribution rose at a more rapid rate than at the high end. The differential was associated with a net improvement in horizontal equity attributable to
‘bringing up the bottom’ of the achievement distribution. In spite of the net gain, however, the annual variation in quintile values offered little evidence of a discernible trend. The quintile analysis provided greater insight into the increased variability observed in the score distribution in 2002. The decrease in equity that year, evident in all of the variance-based measures of horizontal equity described previously, was associated with an increase in the mean score of the top quintile and a decrease in the mean score of the bottom quintile, suggesting that the increased variability in test scores in 2002 occurred across the entire distribution.

The analysis turned next to the three distributional measures of horizontal equity, the Gini coefficient ($G$), the Theil statistic ($T$), and the Atkinson index ($A$). The statistics were computed and compared with benchmark values generated by simulating the annual score distributions under the assumption that all students met the minimum cut-score standard on the test. With the exception of the 2002 results, the overall trend of the distributional measures for the period appeared to have been uneven but seem to have suggested a weak, positive trend. The weakness of the trend, however, must be emphasized due both to the unevenness in year-to-year shifts, and also to the facts that the 1998 data include a smaller $n$ and the 2002 data were associated with a marked, statewide increase in variability in that year. With the exception of 2002 the observed values appeared to converge with the benchmark values ($G', A', T'$), again implying a weak trend toward improvement.
5.2.2.5 Summary of results for Research Question 2

Evidence regarding the second research hypothesis was mixed. The point measures proved to be of little value due to the extremely limited basis of their information. The variance-based measures provided results consistent among the group, with the SDL revealing somewhat greater sensitivity to changes at the low end of the distribution. The general indication of these measures was of a very modest improvement trend. The distributional measures suggested net gains in equity, but of a very modest degree and with, at best, a very weak general trend. Results for the distributional measures did yield evidence that equity was improving in terms of convergence with the minimum-standard equity benchmark simulated for each year. The 2002 results for all of the variance-based and distributional measures departed sharply from the values computed for the other five years. As suggested in the descriptive statistics and initial analyses, results of the test form administered in 2002 might have confounded the analysis by yielding uncharacteristically high variability.

5.2.3 Research Question 3: Did reading scores converge within- and between-schools?

Following the analysis of overall district results, the study examined the separate trends in reading proficiency within and between schools using the factorial decomposition of the Atkinson index and the additive decomposition of the Theil statistic. Evidence was evaluated with respect to whether the observed improvement in district wide equity was associated with changes in the relative shares of within- and between-schools equity.
5.2.3.1 Descriptive statistics

Descriptive statistics by school suggested greater equity between schools than within schools. Variance within schools from year to year did not appear to follow a discernable pattern. All of the school means tended to increase with some exhibiting a more consistent trend of improvement than others. The descriptive statistics raised concerns regarding unequal student numbers for schools over the years, with some schools revealing large percentage gains and losses from year to year. Review of these results with the district administration suggested that the differences were likely due to several factors, including annual volatility in student enrollment, annual volatility in busing and neighborhood attendance zones, districtwide mobility of the population, and incomplete student records due to incomplete records transfers between elementary and middle schools. Results of one-way ANOVA of reading scale scores by school, run for each of the six years in the study period offered evidence regarding changes from year to year in within-schools variance, based on the within-schools mean square term, and year to year changes in between-schools variance, based on the between-schools mean square term. The results showed a net decline in within-schools variance and a net increase in between-schools variance. These findings implied that within-schools equity increased while between-schools equity decreased over the study period.

5.2.3.2 Summary of results for Research Question 3

Overall, no strong pattern was evident for either the within- or between-schools components. Changes in between-schools equity appeared to vary only slightly from year to year along a generally flat line. Consistent with the indications of the descriptive
statistics, the results of the Atkinson decomposition, with the inequality aversion factor $e$ set at $e=.5$, suggested that, with the exception of 2002, within-schools equity appeared to follow a weak, positive general trend of improvement similar to the district results for $G$, $A$, and $T$. Due to the factorial nature of the Atkinson decomposition, the results did not offer any further indication with respect to the changing proportional relationship of within- and between-schools equity over the study period. The Theil decomposition yielded results similar to those of the Atkinson. The additive character of the decomposition permitted direct assessment of the changing proportions of within- and between-schools equity components. Results suggested that the within-schools equity accounted for an increasing share of the total, rising from 94.8% to 95.7% of district equity in reading scores.
CHAPTER 6
CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

A basic prerequisite for the successful maintenance of a democratic form of government is the cultivation of an educated electorate. We refer to this requirement as the democratic imperative for education, imperative because the life of the republic depends upon it. The democratic imperative implies a certain degree of egalitarianism in the sense that all citizens must share at least a minimum level of educational attainment. This egalitarianism is not absolute; any citizen, whether by effort or ability, might acquire more education than is required for participation in the democratic milieu. In this regard the democratic imperative implies what we refer to as egalitarian inequality, a concept referring to the assurance of a minimum standard of education while accommodating, even encouraging, achievement above the minimum.

In the review of literature we traced the conceptual origins of the democratic imperative for education, from the perspectives of John Locke’s political philosophy through Adam Smith’s economic theory, and through the writing of a cross section of early republican leaders, including Thomas Paine, Noah Webster, Thomas Jefferson, and Benjamin Rush, whose diverse political perspectives shared the common theme of giving
voice to the democratic imperative for education and the risks that would attend if institutions were not established to fulfill the requirements entailed in that imperative.

For two hundred years concerns with educational equity were discussed in terms of equity in achievement but were pursued in terms of equity in resources. From Benjamin Rush’s critique of wealth-based disparities in education in the 1780’s, Horace Mann’s efforts to build a system of common schools across the expanding nation, George Strayer’s pioneering contributions to the systematic measurement of educational equity, and continuing through the Civil Rights litigation of the second half of the twentieth century, educational stakeholders and policy makers have struggled to interpret equitable educational achievement by examining the equitable distribution and access to educational resources. Accordingly, research on the measurement of educational equity crystallized into an input paradigm in which the center of interest was educational resources. To the extent that educational achievement was considered, the focus was on the relationship between achievement and resources vis à vis productivity and efficiency problems.

In the wake of a series of judicial and legislative successes, including guarantees of equal access to education for students based on race, sex, ability, economic status, and English proficiency, the limits of achieving educational equalization through resource equalization became evident. Policymakers could ensure student access to schools and classrooms but could not ensure fulfillment of the democratic imperative for education as long as the equity of the educational system was measured in terms of the resources going into the system rather than the student achievement coming out of the system.
Called for as early as the 1840’s by Horace Mann and Samuel Gridley Howe, the development of common, minimum standards of student achievement was undertaken at the state level following promulgation of evaluation criteria for educational programs funded under the *Elementary and Secondary Education Act of 1965*. Adopted first in Florida in 1967, the majority of states had adopted some form of minimum competency testing by the end of the 1970’s. The momentum increased following the 1983 publication of *A Nation At Risk*. The loosely developed minimum competency tests gave way to integrated standards-based reform, leading in turn to the emergence of state adopted content standards combined with policies of performance accountability based on the results of standards-based-state tests. Through the 1990’s, the performance accountability movement steadily gained momentum in education policy, emerging at last as a national mandate for school accountability as part of NCLB.

The emergence of state-based accountability systems, which are built around the articulation and adoption of policies mandating the public expectations for the specific core content and minimum standards of student attainment based on standardized measures of educational achievement, enable the development of a new paradigm in the study and measurement of educational equity, a paradigm we refer to as achievement equity. The present study used as a starting point in this new path of inquiry a set of equity measures and techniques developed in the economics literature and adopted elsewhere in studies of educational resource equity. We have offered an empirical introduction to the measurement of achievement equity by studying the results of sixth grade reading achievement in a large, Midwestern school district.
6.2 Discussion

This provides a discussion of the findings from the study. The section is divided into five parts, covering the equity measures, the decompositions, the study district, the proficiency test, and generalizability.

6.2.1 The equity measures

The study presented and demonstrated four sets of equity measures. The basic functional purpose of an equity measure is to enable the ranking of data distributions. In this study, a central question was whether the distribution of students’ sixth grade reading proficiency scores increased and whether the distribution of scores tended to become more compact, particularly in terms of scores among lower achieving students converging toward the cut score. Addressing this empirical question, we reviewed a broad cross-section of equity measures: point measures, variance-based measures, distributional measures, and decomposable measures. As the literature suggested (Dalton, 1920; Berne & Stieffel, 1984; Sen, 1997), the point measures of horizontal equity are problematic due to the fact that these measures attempt to characterize an entire distribution based on two single data points. In the study, the point measures yielded information generally consistent with the more comprehensive measures but this can be understood to have been coincidental since we might expect that in any given case the behavior of the distributions at these selected points may or may not coincide with the behavior of the rest of the distribution. In spite of the consistency of the results of the point measures in our case, the functional logic of these measures, and the uncertain
relationship they bear to a distribution as a whole, advises against their use as general
measures of achievement equity.

The variance-based measures of horizontal achievement equity, i.e., variance, coefficient of variation, and standard deviation of logarithms, offered generally comparable results. The standard deviation of logarithms offered the unique characteristic of enhanced sensitivity to changes among scores at the lower end of the distribution. This feature did not yield remarkably different results from the other measures, but the $SDL$ earns distinction given the policy emphasis on raising student scores to the cut score.

The distributional measures of horizontal achievement equity, i.e., the Gini coefficient, Atkinson index, and Theil statistic, produced results not remarkably different from those of the variance-based measures. The general pattern of results and magnitudes of differences were generally consistent from measure to measure. The Theil statistic is distinctive here insofar as improvements are indicated by declining values whereas the other two measures improve as values go up. If the analyst were interested in presenting results in a consistent direction, the Theil values might simply be subtracted from zero. From a policy perspective, the Theil statistic is most consonant with the objectives of raising student achievement both among students at the low end of the distribution, but also more generally raising the performance of all students. With the Theil statistic, raising performance at the low end of the student score distribution is exponentially more desirable than trying to improve equality by lowering scores from the high end of the distribution. Additionally, the Theil statistic will indicate an improvement in equity in cases where the entire distribution is shifted upward, i.e., equity
improves when all students improve. These features recommend the Theil statistic in particular for applications in the measurement of achievement equity.

A caveat regarding computational complexity and accessibility must be offered. To a nontechnical audience, the point measures will be particularly attractive for their straightforward simplicity. The variance can, with some effort, be explained to most general audiences but the coefficient of variation and standard deviation of logarithms will go over the heads of many people. The Gini coefficient has the benefit of a simple diagrammatic illustration, the Lorenz curve (Lorenz, 1905). The Atkinson index, in spite of its embedded social welfare function (Sen, 1997), is both computationally and conceptually complex. The Theil statistic, though computationally complex, has the conceptual benefit of bearing an analogous relationship to entropy (Theil, 1967; Sen, 1997).

An alternative approach to the computational/conceptual dilemma is to consider the measures as indicators of ‘sameness’ of student achievement. From this perspective, the Gini and Atkinson measures have the benefit of operating with the confined range of zero to one. Our study has shown, however, that this is somewhat of a Pyrrhic victory insofar as the observed values for the measures had to run to six decimal places, a feature that few policymakers would likely consider intuitive. It must be noted, however, that this phenomenon of extremely small values is peculiar to this study and may be related to the comparatively large numbers of observations. In studies of national incomes (e.g., Conceição & Ferreira, 2000; Litchfield, 1999) and studies of school district expenditures (e.g., Odden, Berne, & Stiefel, 1979; Berne & Stiefel, 1984), the values of the distributional measures can be expressed using two digits. This suggests the possibility
that studies of inter-district achievement equality, with similar \( n \)’s as these other studies, might yield similar results.

Regardless of the measure used, a critical consideration is that the distributional measures offer much more comprehensive information regarding the nature of change in the distribution of student scores than the currently used pass rates can ever tell us. The pass rates are essentially another form of point measure, indicating the proportion of students that fall above or below a particular point in a distribution. Changing pass rates do not inform us regarding changes across the entire distribution. Most importantly, changing pass rates offer no insight into the changing performance of students at the low end of the distribution. Also, pass rates do not offer any indication of whether high performing students are moving farther ahead, standing still, or falling back. For these reasons, the pass rate risks being not simply uninformative but downright misleading. If students near the cut score improve, the pass rate will go up. If students well below the cut score fall back, the pass rate will remain unchanged but equity will decrease and the problem of fulfilling our objective of educating all students will become more remote. The pass rate must be understood for what it offers: an indicator of the proportion of students meeting the standard. In order to understand changes in the distribution of students we must rely on broader measures such as the Theil statistic that can offer relevant information.

From a policy perspective, the analysis and preceding discussion offer insight into two practical questions regarding the use of the equity measures. First, do the equity measures offer important additional information beyond the pass rates? Second, do the equity measures stand on their own as indicators of changes in overall student
achievement? To the first question, the answer is definitely yes. The pass rates for this district revealed no discernible pattern of change over the study period while the equity measures suggested a weakly positive linear trend. To the second question, the analysis revealed that a measure such as the Theil statistic will capture the key information sought for in achievement accountability: rising test scores with particular attention to improvement at the lower end of the score distribution. The analysis also suggests, however, that no single statistic will capture all of the information about a given distribution of scores over time. A prudent policy approach, therefore, would incorporate several measures to track change, and the nature of that change, over time. A combination of pass rates, quintile means, and the Theil statistic would offer a comprehensive battery of equity and achievement indicators.

6.2.2 Decompositions

The most exciting contribution of this study is the decomposition of the achievement equity measures. The decomposability of the Theil statistic, in particular, was shown to offer a means of unfolding achievement equity across more than one level. Here we examined the within- and between-schools levels. The economics literature has demonstrated how these decompositions can be extended to multiple levels (e.g., Akita, 2003; Conceição & Ferreira, 2000; Litchfield, 1999; Sen, 1997) and the present study has demonstrated that this work is transferable to education.
6.2.3 The study district

The study examined changes in students’ sixth-grade reading achievement over a six-year period for a single, large, Midwestern school district. Given the expectation of the state accountability system and, more recently, the reinforcement offered by the NCLB, we would have expected to have observed measurable increases in student achievement from year to year with associated improvements in equity. What we observed appeared to be negligible change that might be considered a general but weak improvement. In terms of meeting the objective of all students passing the test, the district made little progress over the study period. The improvement the district did make appeared to have been concentrated to a certain extent in the low end of the distribution, which is good news from the perspective of the accountability requirements and from the perspective of the underlying democratic imperative for educating all students. In order for the district and the district’s stakeholders to understand more clearly the changes that are taking place, the results of this study suggest that the district, and the state for that matter, should consider approaches to summarizing student achievement on both a distributional and segmented (e.g., quintile) basis. The bad news is that the distribution of student scores, particularly at the low end, does not appear to be moving nearly quickly enough to fulfill the state and federal accountability mandates. A key question for the district is whether attention is being focused on raising the achievement of all students or whether efforts are being adequately orchestrated to address the needs of those students who are below, or far below, the cut score. A key question for policymakers is whether resources have been, or can be, appropriately allocated to meet students’ needs, whether the timeframe for accountability is appropriate, and whether,
more basically, any evidence exists to support the expectation that all students can reach the achievement goals embodied in the tests. Assuming that this district has been exerting efforts to raise student achievement, the present study would not support either of the latter expectations.

Several general problems were discussed regarding the district data. First, the need for accurate, long-term data management is a critical concern. While some records for 1998 were missing, records for the other years appear to be complete. At the school level, the volatility of student numbers was explained in several ways, but this volatility increases the need for consistent strategies for record keeping within and between schools. Students leave one school and go to another but the records may not necessarily follow. The slow, uneven nature of the changing character of student scores over time advises the need for maintaining a data system that enables long term studies. Schools have made impressive progress since the days of paper and pencil records, but the evidence here suggests that improvements and refinements are still necessary. Finally, the particular characteristics of this district, the fact that it is a large, suburban district with higher than average wealth and expenditures and a much higher than average level of teacher salaries, suggests the need to examine patterns of performance across other types of districts to compare both the degree of, and magnitude of change in, equity over time. Further suggestions for research are offered in a later section.

6.2.4 The instrument

The achievement test data used in the study raise several concerns and cautions. The 2002 data, in particular, raise concerns regarding the nature of the test. The
markedly increased statewide SEM for the reading test in that year begs the question of
test stability: how can a well equated form yield such different results? This is not a
simple technical question, but a policy and educational issue regarding the interpretation
of the test scores. What do changes in the distributions of scores mean if the test itself
varies? This problem is related to a second concern about the instrument. The test score
conversion from raw scores to scaled scores expands as scores move out from the cut
score. A one-answer difference between two students makes only a small difference near
the cut score but makes a substantial difference farther from the cut score. The problem
this raises is variability at the extremes. Since we are particularly interested in
determining whether improvements are taking place at the low end of the performance
scale, such increased score volatility in the instrument renders our analyses less stable
and any policy decisions based on those data less tenable. In order to ensure reliable and
valid data for the range of policy making and educational decision making implied in the
accountability laws, the tests must be constructed in such a way as to provide information
for those purposes. Given that the tests were originally designed to determine
proficiency, volatility at the extremes would not have been a concern, but now that we
are trying to understand student achievement across the entire spectrum, both in order to
determine what kind of progress we are making but also to determine the exact nature of
students’ needs, the tests may need to be redesigned.

As expensive and time consuming as that would be, the leading question must be,
‘what do we need to know and what will it take to find out?’ In the mean time, the state
should consider two short term strategies: first, determine the range of appropriate uses
for the score data and make clear recommendations for test use and interpretation;
second, carefully evaluate annual data against multiple years, perhaps reporting only rolling averages in order to mute test-based score volatility.

6.2.5 Generalizability

This study examined a single district. The district was large and diverse in a relative sense, and included twelve differently situated elementary schools. Yet this was only one district, which raises the question of how the results here might compare to other places, contexts, and times. The study offers two major benefits to the continuation of the study of achievement equity. First, we have demonstrated the applicability of equity measures to achievement data and have isolated some of the characteristics of the measures that seem to be most appropriate to the achievement context. Second, we have provided an empirical example that will offer a point of reference for future studies of other districts, other test content, and other grade levels. The results for this district are suggestive, perhaps, that performance may be moving more slowly than the accountability standards might encourage us to hope, but results from a single district are not nearly enough to offer such a conclusion for any but this one district.

6.3 Limitations

The analysis of achievement equity assumed the validity of the proficiency test data. The review of literature discussed ways in which the development process of the test conformed to recommended practice. The review also identified a series of validation studies conducted by various researchers. As Messick (1984) emphasized, validation evidence is never confirmatory of validity but rather of an absence of
disconfirmatory evidence of a particular type. We assume the scores represent a measure of student reading achievement but know only that the scores have so far conformed to the expectations of certain research studies. The question of validity is left open, therefore, and the reader is advised to treat critically both the interpretation of student scores and, ultimately, the interpretation of the measures of achievement equity.

A limitation related to score validity is the assumption of form equality. Each year the Department of Education develops and administers a different form of the reading test. While this has the benefit of reducing teacher and student familiarity with particular items, it introduces potential risks in the comparison of year-to-year scores. Forms are equated following accepted methods of standardized assessment construction but the equating is always within a specified range of tolerance of equality rather than precise equality. This means each form is at least slightly different from the others. This characteristic may have been responsible for the increased level of variation in student scores observed in 2002. Such volatility raises concerns regarding score interpretation, application of the accountability standards, and interpretability of the equity statistics. In this case the statistics were run both with and without the 2002 data in order to evaluate both scenarios. The question that lingers, however, is whether the scores should or should not be considered: the scores are used in the state’s accountability system. Does this imply that the scores should be used in the equity analysis – or should the equity researcher reserve the decision of including or excluding data based on form performance? Including the scores may introduce error and lead to spurious conclusions. Excluding the data puts the researcher in the position of controlling the outcomes by manipulating the selection of data. Neither choice is desirable. Providing side-by-side
analyses might be the most responsible approach, empowering the reader with the greatest degree of information.

The study relied on only one of the state’s five proficiency tests. The use of the sixth grade reading test offered a key measure of competency from the perspective of the democratic imperative for education, recognizing that literacy is a fundamental skill in voting and the acquisition of knowledge regarding the condition of the state. In the context of the state system of standards based accountability, the tests in mathematics, citizenship, science and writing are equally as important as reading. The interpretations made here regarding the apparent improvement trends in achievement equity in the study district are limited to reading achievement alone, and to reading achievement at the grade six level in particular. It is possible that the pattern of achievement equity differed among the other tests and among the other grade levels.

Relying on data from a single district limits the interpretability and generalizability of the results. The results derived in this study may or may not reflect results for this district on other tests nor should we necessarily expect the pattern of reading achievement equity observed in this district to be representative of conditions in other districts. Rather than generalizing results, the study provides the opportunity to generalize the techniques, leaving as an extension of this work the question of the generalizability of results.

The violations of assumptions of ANOVA caution against over-reliance on the findings for the omnibus $F$ test, the post hoc tests of paired comparisons and the post hoc test of linear trend. The alpha adjustments offer some degree of confidence with regard to the statistical violations but no adjustment or manipulation of data can overcome the
absence of control over threats to non-independence of scores due, in part, to the absence of randomization in the assignment of students across schools. As previously discussed, the potential for contamination due to such factors as teacher’s instructional methods, test administration practices, student homework practices, and any other systematic differences among student experiences may undermine the interpretability of the scores by introducing uncontrolled, systematic variance. These factors point to a basic limitation in the state system of standards based accountability, in which standards are promulgated and tests administered but no control measures of student background, teacher and classroom characteristics are collected that might otherwise inform the degree of score interdependence or permit statistical control of such interdependence.

The literature on equity measurement is characterized by an absence of guidance on statistical significance of differences among equity scores. This limitation is due presumably to the fact that economists generally use population income data, which might be assumed to be an error-free estimate of income. Whether such an assumption can be made is left to the interested researcher, the problem introduced here is the fact that student test scores are assumed to be estimates of student achievement with known, or at least measurable, error variance. A statistic of reading achievement equity cannot be taken at face value, nor can the difference between two equity scores; the equity statistics are subject to the same variability as the underlying test scores on which the equity measurement is based. This means that the equity statistics for achievement equity are estimates rather than parameters and must be treated as such. This study offered a preliminary solution to this problem by devising an $F$-test of significance between achievement equity scores. The appropriateness of this test must be examined cautiously
to ensure its functionality. Other statistical tests may be available or be developed that provide greater accuracy.

Similarly, the literature lacks guidance on the interpretation of substantive significance in changes in equity. Evaluation of the substantive nature of observed changes was based in this study on convergence with the minimum achievement equity benchmark represented by all students meeting or exceeding the minimum standard for reading achievement. The substantive question was also considered from the perspective of effect size in the estimation of linear trend in equity statistics across years. The latter approach must be treated with due caution in light of the explanation, above, of interpretations of statistical significance and violations of the assumption of the independence of scores.

6.4 Theoretical, practical, and research implications

This section examines the implications of this study, both in terms of theoretical considerations such as the democratic imperative for education, practical implications such as the concerns raised regarding the inadequacy of proficiency pass rates to convey information about changes in the overall distribution of student scores, and research implications regarding the many questions and issues raised in this paper.

6.4.1 Theoretical implications

This study has provided a conceptual framework for the measurement of achievement equity that integrates the concepts of the democratic imperative for education, egalitarian inequality, and standards-based accountability. Drawing on the
works of theorists John Locke and Adam Smith, and the essays and letters of republican leaders Thomas Paine, Noah Webster, Thomas Jefferson, and Benjamin Rush, we have offered a rationale for the measurement of educational equity outlined in terms of a minimum required level of educational achievement to ensure the maintenance of our republican institutions. In accordance with democratic principles, the articulation of that minimum is left to the will of the people through their elected representatives. Adapting from Pozner (1992), the concept of an achievement minimum implied by the democratic imperative for education led us to the concept of egalitarian inequality, referring to the condition of a community in which all citizens are ensured a minimum education sufficient to enable basic political participation while leaving to individuals, depending on talent, motivation, and/or means, the choice of pursuing achievement beyond the minimum. This combination of ideas, the democratic imperative and egalitarian inequality, have gained a new level of practical meaning and relevance in the wake of the standards-based reform movement, state-based initiatives that were federalized in the NCLB, which mandates state systems of academic content standards, integrated assessments of student proficiency designed to measure achievement with respect to those standards, and accountability policies that hold districts and schools accountable for meeting a specified minimum level of achievement for each of several student subgroups at the school level.

We have argued that standards based accountability is the logical extension of the proposition of achievement equity underlying the Civil Rights litigation for educational equity in cases spanning the second half of the twentieth century (e.g., Brown, 1954; Cannon, 1979; DeRolph, 1997; Lau, 1974; PARC, 1972; Rose, 1989; Serrano, 1971).
Although the remedies tended to focus on the equalization of educational resources and services, the ultimate goal was achievement equity vis à vis the democratic imperative. Standards based accountability departed from traditional equity remedies by directly establishing equity in educational achievement rather than attempting as plaintiffs had in their equity lawsuits to promote achievement equity indirectly by manipulating the distribution of educational resources. This transformation in public policy, from a resource orientation to an achievement orientation, has offered a rationale for a corresponding reorientation of the measurement of educational equity toward student achievement. Since, in accordance with the democratic imperative for education, the immediate objective of the educational system is student achievement, the appropriate focus of equity measurement is achievement equity. Since the achievement goals represent what has been defined as the adequate minimum determined through democratic processes, the measurement of achievement equity must accommodate the concept of egalitarian inequality. Since states’ content standards and assessments represent the definition of the democratic minimum adopted by our republican representatives, the measurement of achievement equity should focus on the distribution of student scores on these assessments. Since the standards specifically require fulfillment of the minimum achievement standards across individual student subgroups, the achievement equity measurement methodology must offer decomposability into subgroup equity components in order to understand the absolute and relative standing of each group.

This study, adopting the approach just described, promises to realign the measurement of educational equity with the goals of educational equity, shifting away
from the question of resource equity to the more fundamental question of achievement equity. This shift represents not simply a change in the unit of measure, scores versus dollars, but a change in the equity paradigm itself, from a conceptual construct in which equity is understood as equality of resources and services to a construct in which educational equity is understood in terms of the set of measurable traits we associate most directly with the educational process, i.e., students’ educational achievement.

### 6.4.2 Practical implications

The proficiency rate is the prevailing method by which analysts and policy makers examine and compare student performance within and between schools, districts, and student groups (e.g., Ohio Department of Education, 2004). Having demonstrated the critical limitations of proficiency rates, this study has offered an empirical illustration of a set of equity measures that provide information on changes in the structure of the distribution of student scores. Grouping the measures into point-, variance-based, and distributional measures of achievement equity, the study offers confirmatory evidence of findings that the point-measures are of no practical value to an achievement equity inquiry and that the standard deviation’s dependence on the mean renders it unsuitable. Of the remaining measures, $CV$, $SDL$, $G$, $A$, and $T$ each bears certain desirable properties that none captures equally well, but that the Theil statistic comes closest by incorporating the set of characteristics most consistent with achievement accountability policy: mean independence; sensitivity to changes throughout the distribution; increased sensitivity to changes at the low end of the distribution, and additive decomposability, which enables separate analysis of subgroup equality in such a way that the sum of the subgroup
components adds to the total, overall equity, a feature particularly well suited to the evaluation of student subgroup equity relative to overall equity.

For educational analysts and policymakers, the measurement of horizontal equity can offer a far more informative approach, as compared with pass rates, to assessing questions of group improvement over time. Determining the proportion passing the test is one key piece of information for accountability. The two immediate follow-up questions of performance accountability are, ‘Is improvement increasing over time?’ and ‘Are scores getting closer together, particularly from the bottom up?’ The measurement of horizontal equity, with its emphasis on measuring all scores in relation to all other scores, directly addresses the latter question and, by virtue of turning attention to score distributions rather than simple cut score bifurcations, provides as a byproduct the score data needed to answer the first question. From both an educational accountability perspective and an educational planning perspective, meaningful answers to these questions, falsely implied by proficiency data, provide policymakers and educators with information needed to understand the nature of changes in score distributions over time, between groups, and within groups.

In order to better understand the implications of observed changes in equity statistics, two particularly relevant concerns are whether the changes are statistically significant or substantively significant. Using score variance and benchmarking against the cut score, this study has provided a tentative approach to the evaluation of significance. While any increase in equity might be a desirable outcome, we must be confident that observed increases are not simply the result of chance variation in student scores. We also need to know whether we are making adequate progress to meet our
equity objectives, such as the goal for 95% of students to reach the performance standards by the year 2014. The systematic evaluation of statistical and substantive significance probed in this study provides an important starting point in this inquiry.

6.4.3 Research implications

This study has opened the path to a new paradigm in the theory and practice of educational equity measurement. The research agenda promises to be wide ranging and complex. Highlights of this research agenda will minimally encompass the following:

1. The study used data collected from a single, district. How will achievement equity analysis unfold at a multi-district, regional, state, or multistate level?

2. The study used only a single measure, reading scores. The number of variables must be increased in order to better reflect the fact that state accountability systems include multiple measures at multiple grades and to begin to understand the interpretation of multiple equity indicators.

3. It is possible that underlying the multiple assessment measures used in a state’s accountability system reside a smaller number of complex constructs, e.g., a general reading competency construct as opposed to 4th, 6th, and 10th grade reading competencies. A thorough understanding of achievement equity would require measurement of equity on such constructs. Alternatively, or additionally, since schools face the need to face multiple achievement objectives simultaneously, a relevant measure of achievement might be a multivariate scale of combining multiple measures.
4. The equity literature, though mature, remains dynamic. This study offered an analysis of a cross section of measures, others must be explored and understood. Additional measures may yet await articulation and development. Two measures not included here are the McLoone index (1965), which examines equity among scores residing below a distribution’s median, and the reverse McLoone (Verstegen, 1996), which examines equity among scores residing above a distribution’s median. Placing these measures within the context of achievement equity suggests the possibility of developing a family of criterion referenced equity measures in which the median is one criterion of interest and the cut scores on the tests would offer other criteria. The number of possible relative (e.g., median) and absolute (e.g., cut score) criteria implies at least two subfamilies of criterion-referenced equity measures.

5. This study introduced the Theil and Atkinson decompositions but limited the inquiry to within- and between-schools equity. The flexibility of the decomposition techniques lends them to the examination of the entire range of subgroup analyses mandated by accountability standards. Equity decompositions at the state level might address regional or typological groupings of school districts (e.g., Appalachian, suburban, agricultural, urban), or program groups (e.g., a group of schools implementing a common reading intervention program). Within school districts, the analysis here addressed within- and between-schools variation in equity. Further geographical breakdowns might include zip code and subdivision groupings in order to examine equity among similar demographic groups or attendance groups.
6. The validity of the state tests must be examined carefully at each grade level. This study revealed a wide change in score variability during a particular year, 2002, suggesting the need to better understand the stability of the instrument and the form equating process. This study also revealed concerns with score stability with increasing distance from the cut score. Given the large proportion of students below the cut score and the need to measure improvement in that region over time, it is vital that researchers, and policymakers, gain a clear understanding of the degree, and implications for accountability and equity, of score instability across the score scale.

7. We have examined achievement equity here in isolation from educational inputs. An examination of comparisons among changes in achievement equity and fiscal equity might help us begin to understand the relationship between the two over time.

8. With increasing interest in questions regarding value-added models of educational achievement, technological advances in tracking student scores over time will enable the measurement of achievement equity over time for a given cohort of students, as distinguished from the strategy in the current study of examining changes in equity for different cohorts at a given grade level over time. Longitudinal cohort studies may also offer insights into the relationships between value-added and equity.
6.5 Reflections

The maintenance of democracy requires the adequate education of our youth. Americans have struggled since the founding of the republic to define the scope of this democratic imperative and to fulfill its requirements. Over the course of the twentieth century, the educational minimum was defined in terms of fiscal resources for education. Educational equality, in turn, has been measured in terms of the equality of the distribution of fiscal resources.

This paper has contended that the shift away from fiscal equity toward achievement equity calls for a shift in our approach to the measurement of educational equality. The study expanded the conceptual framework for measuring educational equality by incorporating the ideas of the democratic imperative for education and standards-based reform. These two ideas were brought together in the concept of egalitarian inequality, a condition marked by all students reaching or exceeding the level of educational attainment necessary to function in democratic society.

Drawing on this framework and established measures of equality, the study offered an empirical analysis of achievement equity. Posing the question of whether achievement equity has improved in recent years, as we would expect based on state and federal accountability legislation, we examined changes in equity in the distribution of reading proficiency scores drawn from a large, Midwestern school district for the years 1998 through 2003. Characteristics of the equality measurement techniques were described and compared. Statistical significance of observed changes in equity values were tested based on observed variability in student achievement. Following the federal mandate for disaggregating student subgroup achievement, decomposition techniques
were applied to determine within- and between-groups contributions to district equity. Results of decomposition measures were presented and compared.

Implications for policy makers and equity analysts were discussed with regard to gaining a more systematic understanding of the achievement gap and how, or whether, schools and districts are making progress toward closing those gaps. Recommendations for future work were offered both for broader, state and national studies and for more detailed district, school, and subgroup decompositions.
BIBLIOGRAPHY


_DeRolph v State (DeRolph I)_ 677 N.E. 2d 733 (1997).

_DeRolph v State (DeRolph II)_ , 78 Ohio St. 3d 193 (2000).

_DeRolph v State (DeRolph III)_ , 93 Ohio St.3d 309 (2001).

_DeRolph v. State (DeRolph IV)_ , 97 Ohio St.3d 434 (2002).


Title IX, Education Amendments (1972) Title 20 U.S.C. Sections 1681-1688.


APPENDIX A

RAW SCORE TO SCALED SCORE CONVERSION TABLE

GRADE 6 READING, MARCH 2003
<table>
<thead>
<tr>
<th>Raw</th>
<th>Scaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>89</td>
</tr>
<tr>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td>4</td>
<td>114</td>
</tr>
<tr>
<td>5</td>
<td>122</td>
</tr>
<tr>
<td>6</td>
<td>129</td>
</tr>
<tr>
<td>7</td>
<td>136</td>
</tr>
<tr>
<td>8</td>
<td>141</td>
</tr>
<tr>
<td>9</td>
<td>146</td>
</tr>
<tr>
<td>10</td>
<td>151</td>
</tr>
<tr>
<td>11</td>
<td>155</td>
</tr>
<tr>
<td>12</td>
<td>159</td>
</tr>
<tr>
<td>13</td>
<td>163</td>
</tr>
<tr>
<td>14</td>
<td>166</td>
</tr>
<tr>
<td>15</td>
<td>170</td>
</tr>
<tr>
<td>16</td>
<td>173</td>
</tr>
<tr>
<td>17</td>
<td>176</td>
</tr>
<tr>
<td>18</td>
<td>179</td>
</tr>
<tr>
<td>19</td>
<td>182</td>
</tr>
<tr>
<td>20</td>
<td>185</td>
</tr>
<tr>
<td>21</td>
<td>187</td>
</tr>
<tr>
<td>22</td>
<td>190</td>
</tr>
<tr>
<td>23</td>
<td>193</td>
</tr>
<tr>
<td>24</td>
<td>195</td>
</tr>
<tr>
<td>25</td>
<td>198</td>
</tr>
<tr>
<td>26</td>
<td>201</td>
</tr>
<tr>
<td>27</td>
<td>203</td>
</tr>
<tr>
<td>28</td>
<td>206</td>
</tr>
<tr>
<td>29</td>
<td>209</td>
</tr>
<tr>
<td>30</td>
<td>212</td>
</tr>
<tr>
<td>31</td>
<td>215</td>
</tr>
<tr>
<td>32</td>
<td>217</td>
</tr>
<tr>
<td>33</td>
<td>222</td>
</tr>
<tr>
<td>34</td>
<td>224</td>
</tr>
<tr>
<td>35</td>
<td>227</td>
</tr>
<tr>
<td>36</td>
<td>230</td>
</tr>
<tr>
<td>37</td>
<td>234</td>
</tr>
<tr>
<td>38</td>
<td>238</td>
</tr>
<tr>
<td>39</td>
<td>242</td>
</tr>
<tr>
<td>40</td>
<td>246</td>
</tr>
</tbody>
</table>

Table A.1: 6th grade reading score conversion table

Continued
Table A.1 continued

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>250</td>
</tr>
<tr>
<td>42</td>
<td>255</td>
</tr>
<tr>
<td>43</td>
<td>261</td>
</tr>
<tr>
<td>44</td>
<td>267</td>
</tr>
<tr>
<td>45</td>
<td>275</td>
</tr>
<tr>
<td>46</td>
<td>283</td>
</tr>
<tr>
<td>47</td>
<td>294</td>
</tr>
<tr>
<td>48</td>
<td>308</td>
</tr>
<tr>
<td>49</td>
<td>332</td>
</tr>
<tr>
<td>50</td>
<td>355</td>
</tr>
<tr>
<td>51</td>
<td>254</td>
</tr>
<tr>
<td>52</td>
<td>258</td>
</tr>
<tr>
<td>53</td>
<td>261</td>
</tr>
<tr>
<td>54</td>
<td>266</td>
</tr>
<tr>
<td>55</td>
<td>271</td>
</tr>
<tr>
<td>56</td>
<td>276</td>
</tr>
<tr>
<td>57</td>
<td>282</td>
</tr>
<tr>
<td>58</td>
<td>290</td>
</tr>
<tr>
<td>59</td>
<td>299</td>
</tr>
<tr>
<td>60</td>
<td>313</td>
</tr>
<tr>
<td>61</td>
<td>336</td>
</tr>
<tr>
<td>62</td>
<td>359</td>
</tr>
</tbody>
</table>