Quantifying Quality:
The Effects of Score Transformation Method and School Demographics on School Rankings Under the Elementary and Secondary Education Act

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Alison Elizabeth Phillips

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Frances Lawrenz, Advisor
Dr. Andrew Zieffler, Co-Advisor

July 2015

Acknowledgements

Successful completion of this dissertation would not have been possible without the input, support, and humor of my colleagues, friends, and family.

From the very beginning of brainstorming viable research questions to engaging in countless conversations at 6:00 a.m. to process my overnight analysis epiphanies, I am deeply indebted to Kat Edwards and her never-ending patience.

I must also thank my parents, Julien and Jeff Phillips, for assuring me that going back to school was not a mistake, for combing through pages and pages of text with the eyes of copyeditors, and for being my most reliable practice audience along the way.

Without question, I would not have made it through the homestretch of writing and defending my dissertation were it not for the daily support of my fiancé and true partner, Scott Johnson. I look forward to years of signing our holiday cards as Dr. and Mr. Phillips.

My acknowledgements could not be complete without mention of my advising team and committee. To Dr. Selcen Guzey and Dr. Michael Harwell, a sincere thank you for your support in increasing the rigor of this work. To Dr. Frances Lawrenz and Dr. Andrew Zieffler, a very special and enthusiastic thank you for all of your academic and emotional support over the past five years.

Lastly, thank you to the leaders in federal government who have worked tirelessly to create, maintain, and revise legislation aimed at holding states responsible for providing the best educational experiences possible to all students. Never stop improving your standards.

Abstract

Reasons for quantifying and ordering relative school achievements as a measure of school quality are numerous. They range from informing parents about where to enroll their children to complying with federal reporting and accountability requirements. Even after accepting the premise that results on state tests designed to measure student mastery of subject standards can serve as a proxy for the measure of a school's quality, questions remain about how individual student results should be transformed into a school-level measure in a way that is more reflective of how a school is serving its students than of what type of students a school serves.

This study examines the effects of using different score transformations from the same test results to rank schools by investigating three questions: (1) What effect does the method of transforming student scores on Minnesota state exams have on relative Minnesota school performance rankings over time?; (2) What effect do school demographics have on relative Minnesota school performance rankings over time?; and (3) What effect do the interactions of method of transforming student scores on Minnesota state exams and school demographics have on relative Minnesota school performance rankings over time?

A unique opportunity for robust analysis of a complete set of statewide individual level testing and enrollment records was available though a special agreement with the Minnesota Deaprtment of Education. Comparison of multilevel models shows that simpler score transformations lead to quantification of quality and relative school

rankings that are the least related to the demographic characteristics of the students a school serves.

Table of Contents

List of Tables

List of Figures

Quantifying Quality:

The Effects of Score Transformation Method and School Demographics on School Rankings Under the Elementary and Secondary Education Act

Chapter 1

Introduction

Attempts to quantitatively measure school quality are common. Purposes of this measurement are numerous, ranging from states complying with the reporting requirements of the federal Elementary and Secondary Education Act (ESEA) to awards of recognition such as the National Blue Ribbon Schools (NBRS) award. However, the ambiguity of the term "quality" combined with the varying agendas of educational organizations, special interest groups, and national, state, and local governing bodies has led to an abundance of inconsistent definitions and measures of quality used to evaluate relative school excellence.

U.S. News releases an annual national ranking of high schools based on three factors: performance levels, measured by student performance on state accountability tests, controlling for relative school poverty; proficiency rates, measured by proficiency on state accountability tests for disadvantaged student subgroups (e.g. ethnically or economically disadvantaged) relative to state averages; and college preparation, measured by participation and performance on Advanced Placement exams and/or International Baccalaureate exams (Morse, 2013).

Other organizations, such as the Specialist Schools and Academies Trust (SSAT), suggest school quality is better measured by the presence and strength of subjective

characteristics. In 2010, the National Head Teacher Steering Group of the SSAT published twelve characteristics that they believed to be linked to school quality. In addition to teaching and student performance, these characteristics include: having strong leadership, developing and encouraging character, and maintaining strong links with businesses and the community (Anonymous, 2010).

1.1 Federal Requirements

Under the ESEA, states are required to measure schools' progress towards having all students meet rigorous academic standards. The Minnesota Department of Education (MDE) currently uses a tool known as the Multiple Measurement Ratings (MMR), a combination of scores in proficiency, growth, achievement gap, and graduation domains, to rank schools and meet federal reporting requirements under the state's approved ESEA waiver (Minnesota Department of Education [MDE], 2012a).

Though whether or not one can really ever accurately measure either relative or absolute school quality through transformations and aggregations of student test scores is a hotly debated issue, the requirement to do so remains a part of federal accountability law. Thus, questions of whether or not school quality should be measured by test scores are of less immediate practical importance than questions regarding what impact the methods used to transform test scores into a quality index for schools have on a school's relative quality ranking.

1.2 Aspects of a Good Ranking System

Accepting the requirements of ESEA in its current form, two aspects of choosing an appropriate methodology to measure school quality via test scores emerge as being

particularly important: (1) how understandable the calculation of the ranking is to the public and (2) how independent of school characteristics, other than "quality," the ranking is. A numeric proxy for quality should be an indicator of attributes indelible to the school, and not an indicator of what types of students the school serves. As such a proxy, relative school rankings should be as independent of the demographics of the students a school serves as possible.

The highly public nature of these rankings presents its own set of challenges. School leaders are expected to understand how test scores are used to determine relative rankings and to set test-related goals to improve their school's standing. It is generally assumed that many parents will use school rankings to inform decisions about where to enroll their children and that local media will use rankings to promote and compare schools and districts. Having a calculation that is as simple as possible supports appropriate interpretations of rankings by parents, schools, and the public at large. Ultimately, figuring out which method of score transformation generates the "best" school ranking requires looking at all of the data together and figuring out the advantages and disadvantages of each kind of transformation (Seife, 2010, p. 26).

Chapter 2

Literature Review

Federal law, grant opportunities, and public demand for school monitoring all require the quantification of school quality. Yet, consistent measures of school quality are not used either across or within these purposes. A need exists for a robust method of determining school quality that uses relevant school data and can be applied to a diverse set of situations calling for measures of school quality or school rankings. To better understand what such a method needs to include, it is vital to have basic knowledge about the requirements of the law and the structure of existing state and national measures of school quality.

2.1 The Elementary and Secondary Education Act

In 1965, President Lyndon B. Johnson changed the national face of educational legislation with his creation of the Elementary and Secondary Education Act (ESEA). While Johnson himself said it was "'the greatest breakthrough in the advance of education since the Constitution was written'" (McKay, 1965, p. 427), others agreed that ESEA would "without question, go into the books as the most significant educational achievement of any Congress in this century, indeed if not in the entire history of the nation" (McKay, 1965, p. 427). Part of Johnson's War on Poverty, the original ESEA (1965) consisted of six Titles, the most notable of which was "Title I–Financial Assistance to Local Educational Agencies for the Education of Children of Low-Income Families." Title I formally took notice of the relationship between economics and educational success and promised financial assistance in the form of grants to aid schools

in enhancing and expanding educational programs aimed at "meeting the needs of educationally deprived children" (ESEA of 1965, 79 Stat. 27).

With reauthorization occurring every three to five years, ESEA has gone through some major changes during the last five decades. The most salient reauthorizations took place in 1994 and 2001 with Bill Clinton's reauthorization known as the Improving America's Schools Act (IASA) and George W. Bush's reauthorization titled the No Child Left Behind Act (NCLB). A summary of these reauthorizations and their impacts appears in Table 1.

2.1.1 Improving America's Schools Act. IASA was passed into law shortly following the Goals 2000: Educate America Act (Goals 2000 Act) in 1994, which represented the first piece of significant educational legislation since the Regan administration (State's Impact, 2009). Goals 2000 was based on legislation introduced by George H. W. Bush in 1991 called America 2000 (State's Impact, 2009), and centered around eight national education goals briefly titled: (1) School Readiness; (2) School Completion; (3) Student Achievement and Citizenship; (4) Teacher Education and Professional Development; (5) Mathematics and Science; (6) Adult Literacy and Lifelong Learning; (7) Safe, Disciplined, and Alcohol- and Drug-Free Schools; and (8) Parental Participation (Goals 2000, 108 Stat. 133). Under Title III: State and Local Education Systemic Improvement of Goals 2000, the law articulated that states applying for funds needed to submit an improvement plan including "a process for developing or adopting State content standards and State student performance standards for all students […]" (Goals 2000, 108 Stat. 162), with the intent that through curricular alignment to these

standards "students' mastery of […] English, mathematics, science (including physics), history, geography, foreign languages, the arts, civics and government, and economics" (Goals 2000, 108 Stat. 162) would improve, thereby meeting the third national goal.

Title I of the IASA, the reauthorization of the ESEA, intensified the focus on state-defined educational standards. Though still aimed at closing the achievement gap between students coming from different economic backgrounds, IASA set forth new state program requirements necessary to apply for grant funds. The first of these requirements that would permanently change the structure of ESEA was the establishment of "challenging standards" (IASA of 1994, 108 Stat. 3523). Any state having developed academic standards under Title III of Goals 2000 would be able to use those same standards when applying for an IASA grant. However, a requirement of the standards was that they would, at a minimum, be established for reading or language arts and mathematics and "include the same knowledge, skills, and levels of performance expected of all children" (IASA of 1994, 108 Stat. 3524). Student performance related to each state's academic standards would be measured by state-designed standards-aligned annual assessments and would need to be reported at four levels: not proficient, partially proficient, proficient, and advanced. Additionally, states would need to define and measure "yearly progress" of schools and local educational agencies (LEAs) towards meeting their student performance standards (IASA of 1994, 108 Stat. 3524). IASA afforded great flexibility to the states while ensuring that all students within schools receiving Title I grant funds, even those students from economically disadvantaged

backgrounds, would be offered equitable educational experiences and would be measured by the same scale.

2.1.2 No Child Left Behind. Another bipartisan piece of legislation, ESEA was reauthorized in 2002 as the No Child Left Behind Act of 2001. Title I of NCLB kept many of the same elements as Title I of IASA: the setting of state standards in at least mathematics and reading or language arts – with the additional requirement that standards be set in science by the 2005-2006 school year – and the implementation of a "single state-wide accountability system" (NCLB of 2001, 115 Stat. 1445) to monitor the progress of schools and LEAs in meeting those standards for all children. New requirements were also introduced. Adequate yearly progress (AYP) was redefined with an ultimate goal of all students being proficient in reading or language arts, mathematics, and science by 2014. To reach this overarching goal, states would need to set annual goals for increasing proficiency among all students and in student subgroups disaggregated by race, special education status, gender, limited English proficiency status, and economic background (State's Impact, 2009). Failure to meet AYP goals in any subgroup would result in a school failing to make AYP overall. Failure to meet AYP for two or more consecutive years would result in increasingly stringent corrective action being taken, culminating in complete restructuring of the school after five years (State's Impact, 2009). NCLB raised the stakes in providing federal aid to states for education by attaching consequences for not producing measurable and equitable educational outcomes across all students.

2.1.3 Minnesota's ESEA waiver. No reauthorization of the ESEA has taken place during the Obama administration, though a blueprint for what such reauthorization might look like was released in early 2010 (U.S. Department of Education [USDOE], 2010). However, states have been granted opportunities for flexibility under the requirements of NCLB through state application for and federal approval of ESEA flexibility waivers. Beginning in 2011, the U.S. Secretary of Education offered states the option of applying for flexibility in ten aspects of the 2002 ESEA reauthorization. In addition to some flexibility in use of funds for different types of schools and requirements of corrective action against schools and LEAs not making AYP, approved states are granted flexibility in determining new annual measurable objectives (AMOs) used in assessing AYP (USDOE, 2012)

In 2012, Minnesota submitted a request for an ESEA flexibility waiver that was approved for use through the end of the 2013-2014 school year. Of the four principles required to be addressed in the flexibility application – college- and career-ready expectations for all students; state-developed differentiated recognition, accountability, and support; supporting effective instruction and leadership; and reducing duplication and unnecessary burden (USDOE, 2012) – the second principle had the most salient impact on the ways in which Minnesota measures relative school quality.

To receive flexibility in state-developed differentiated recognition, accountability and support, a state's submitted ESEA waiver application needed to outline a system that would, in part, use measurements of student achievement, student growth, and achievement gap reduction to create ordered list(s) of schools so that annually a

Table 1.
Summary of laws affecting student assessment and measurement of school quality

Year	Act	Major implications
1965	Elementary and Secondary Education Act (ESEA)	Title I makes financial assistance available to schools to serve students from economically disadvantaged backgrounds.
1994	Improving America's Schools Act (IASA)	New requirements of academic standards in both reading and math, and measurement of students' progress towards meeting those standards are introduced.
2002	No Child Left Behind (NCLB)	Consequences for not achieving Adequate Yearly Progress (AYP) towards 100% of students demonstrating proficiency in standards are added to Title I.
2011	ESEA Waivers	States are allowed to create their own systems of differentiated recognition, accountability, and support. These systems must use measures of achievement, student growth, and achievement gap reduction to order schools.

percentage of the top schools receiving Title I funds under NCLB could be identified as "reward schools," and at least 5% of the bottom schools receiving Title I funds could be identified as "priority schools" (USDOE, 2012).

Minnesota's waiver specifies that the top 15% of schools identified through the state's Multiple Measurements Rating (MMR) be designated Reward, the next 25% of schools – those ranking between the top 15% and 40% – be designated Celebration Eligible, schools in the bottom 5% be designated Priority, and schools in the bottom 25%, that are not already designated Priority by the MMR, or Focus by a separate focus rating system, be designated Continuous Improvement (MDE, 2012a).

With school ranking systems such as the MMR, which is based on a weighted average of index scores across the domains of proficiency, growth, achievement gaps, and graduation, it is important that the ways in which achievement, growth, and achievement gaps are calculated be consistent and meaningful. There exists a dearth of referenced research behind the wide variety of methods for quantifying and aggregating achievement utilized by the Minnesota Department of Education to create ordered lists of relative school quality. It is left to the local or global user to determine the internal and longitudinal reliability of each method.

2.2 Achievement

In Minnesota, students' progress toward meeting the state's grade-level academic standards in mathematics and reading is measured by the Minnesota Comprehensive Assessments series III (MCA-III) or one of the alternate assessments – the MCA-Modified (MCA-M or MOD) or the Minnesota Test of Academic Skills (MTAS) – in grades 3-8 (grades 5-8 for MCA-M), as well as in grade 10 for reading and grade 11 for mathematics. To examine how achievement can be reported, it is necessary to have a cursory understanding of the scoring of these tests.

2.2.1 Test scoring. For each of these three accountability tests raw scores are converted to scale scores. When standards within a subject stay the same over consecutive years, scale score comparisons across years are valid within test, grade, and subject (MDE, 2012b). However, reading standards assessed changed in the 2013 administration of the tests for all grades and math standards assessed changed in 2011 for

grades 3-8 and in 2014 for grade 11, making test results not comparable within grade and subject to earlier years.

Another complication in interpreting scale scores is the tests are on different scales. The MCA-III scale scores range from G01 to G99, where G is the grade in which the test was administered. Ranges of MCA-M and MTAS scale scores differ by subject and grade but are all somewhere between 0 and 325. For illustrative purposes, a summary of scale score ranges, means, and standard deviations from the 2013 test administration is given in Table 2.

Further, because the tests are designed to measure student performance on grade-level standards (MDE, 2012b), and these standards are not cumulative across increasing grades, scale scores cannot be directly compared across grades within a year or across years for an individual student who has made expected grade progression. That is, it is not possible to make definitive statements about the relative performance of two students in different grades. For example, if a grade three student earns a scale score of 355 and a grade five student earns a scale score of 560 on the MCA-III reading exam, it is not necessarily correct to say the grade five student did "better" than the grade three student as each is being assessed on different standards. Similarly, it is not possible to say with certainty that a student who earned a scale score of 625 on her MCA-III math test in 2012 and a scale score of 735 on her MCA-III math test in 2013 is improving. The use of different scales for the 38 different tests combined with the different distributions of scale scores within each test makes aggregation using means of scale scores across tests and grades to the school or district level nonsensical. The different scales also make value-

Table 2.
Scale score ranges, means, and standard deviations

Test name	Subject	Grade	Range	μ	σ	N
MCA-III	Reading	3	301-399	351.7	20.3	62452
		4	411-490	450.2	15.4	61233
		5	517-591	554.1	14.7	58697
		6	606-699	652.9	17.5	60001
		7	703-798	750.4	17.8	60695
		8	802-898	850.1	17.6	60254
		10	1013-1094	1053.0	14.7	61891
MOD-III	Reading	5	150-286	191.7	15.2	1413
		6	146-286	194.3	16.9	1904
		7	128-314	196.1	22.3	1938
		8	153-291	194.8	15.6	1870
		10	141-281	199.2	20.8	1248
MTAS-III	Reading	3	58-268	204.3	38.4	826
		4	36-258	209.7	42.8	941
		5	84-249	205.9	30.2	954
		6	107-236	205.7	23.1	918
		7	124-223	200.7	16.3	933
		8	100-235	204.8	25.2	852
		10	150-237	205.2	28.2	807
MCA-III	Math	3	315-399	357.1	15.6	62773
		4	409-499	457.8	17.6	61464
		5	515-586	551.7	12.9	59007
		6	611-688	650.8	13.5	60144
		7	718-782	750.2	11.4	60574
		8	813-888	851.3	13.7	59911
MCA-II	Math	11	1101-1199	1148.5	19.5	61029
MOD-III	Math	5	153-236	188.7	11.2	1320
		6	149-251	187.6	9.7	1859
		7	165-234	188.9	7.7	2146
		8	167-232	190.0	7.6	2215
MOD-II	Math	11	165-233	189.2	8.2	1343
MTAS-III	Math	3	103-257	205.1	30.2	798
		4	96-251	202.7	26.2	917
		5	142-225	198.3	13.4	919
		6	140-226	199.8	12.1	920
		7	124-229	196.7	15.1	926
		8	119-237	198.9	17.1	873
MTAS	Math	11	159-216	197.6	9.3	809

added measures of student achievement or school quality very difficult to calculate or model as such models require the difference in a student's current performance and previous performance to have defined meaning (Hanushek & Rivkin, 2010; Harris, 2011).

2.2.2 Defining proficiency. Harkening back to the four levels of reported proficiency first required by IASA, cut scores are determined for the MCA-III, MCA-M, and MTAS that split the scale score range into the four achievement levels of Does Not Meet the Standards (D), Partially Meets the Standards (P), Meets the Standards (M), and Exceeds the Standards (E). For the MCA-III, the cut scores for P and M are G40 and G50, respectively, regardless of the grade, subject, or year. Similarly, for both the MCA-M and MTAS, the cut scores for P and M are 190 and 200, respectively. For each of the three tests, in each subject and each grade, the cut score for E is determined separately (MDE, 2012b). As with scale scores, the school-level proficiency rates differ greatly by grade, subject and test. For illustrative purposes, a summary of percent proficiency means and standard deviations from the 2013 test administration is given in Table 3.

2.2.3 School achievement. While individual student test results are used to assess student mastery of standards and to help teachers and school administrators identify strengths and weaknesses in their teaching and/or curriculum, test results aggregated by the eight population subgroups (American Indian, Asian/Pacific Islander, Black, Hispanic, White, Free/Reduced Price Lunch, Limited English Proficiency, and Special Education), and at the school and district level are also common. Such aggregate level data are used in the MMR to assess the extent to which schools and districts are equally

Table 3.
Percent proficiency means and standard deviations

Test name	Subject	Grade	μ	σ	N (schools)
MCA-III	Reading	3	51.7	22.0	925
		4	49.1	20.9	918
		5	57.8	22.2	901
		6	49.5	24.2	741
		7	39.7	24.5	691
		8	40.9	26.1	720
		10	44.1	27.1	791
MOD-III	Reading	5	32.5	36.7	475
		6	38.5	35.5	442
		7	45.7	36.5	385
		8	35.5	32.3	375
		10	52.0	36.6	282
MTAS-III	Reading	3	72.1	38.1	383
		4	70.2	37.6	429
		5	71.0	38.6	434
		6	71.7	36.5	331
		7	66.9	38.4	317
		8	68.6	36.8	301
		10	72.3	35.7	251
MCA-III	Math	3	65.2	23.8	926
		4	64.7	24.3	919
		5	52.6	23.8	903
		6	45.3	26.4	741
		7	39.5	26.3	687
		8	41.0	28.2	718
MCA-II	Math	11	29.8	25.7	793
MOD-III	Math	5	18.6	30.7	468
		6	13.7	25.0	419
		7	8.1	18.5	385
		8	11.4	21.7	396
MOD-II	Math	11	9.0	19.4	303
MTAS-III	Math	3	70.3	38.9	382
		4	75.8	36.2	428
		5	64.0	40.8	423
		6	75.4	33.5	334
		7	49.8	39.1	315
		8	68.5	38.3	304
MTAS	Math	11	49.5	40.6	264

serving all students and to create the ordered lists (separated by four types of school based on grades served: elementary, middle, high, and other) from which Reward, Celebration Eligible, Continuous Improvement, and Priority schools are identified (MDE 2012a).

Quantifying subgroup- or school-level achievement can be thought of as analogous to computing a composite score for an individual across a battery of tests. But, as in the case of composite scores, the method for transforming individual test scores in preparation for aggregation to the subgroup- or school-level must be chosen carefully so as to make the aggregate value as meaningful as possible. Since every school in the state that serves students in at least one grade between grade three and grade eleven will have achievement information from some subset of the 38 tests given, a meaningful aggregate value will not reward or penalize a school based on the inclusion or exclusion of any test. The score transformations used to create a relative school ranking need to account for differences in the tests given to make school comparisons valid.

2.3 Methods of Score Transformation

Score transformations that might provide control for the different data inputs different schools will have are already in use within education. Several distinct methods of score transformation are already in use by the Minnesota Department of Education (MDE) to order schools according to relative performance or quality. National and international assessment results from studies such as the Trends in International Mathematics and Science Study (TIMSS), the Program for International Student Assessment (PISA), and the National Assessment of Educational Progress (NAEP) are

also used to order states and/or countries according to relative performance. Aggregation methods used in such large-scale assessments may be applicable to statewide assessment results, though currently they are only used within grade and test subject. The dual test outcomes of scale scores and achievement levels as well as the between-subject, between-form, and between-grade test differences make it possible for tests results to be transformed and aggregated to the school-level in numerous ways. There is no indication that the relative school rankings produced by these various transformations have been evaluated for reliability or compared to one another for consistency.

2.3.1 Percent proficiency. Perhaps the most widely used and easily understood method of aggregate-level test score reporting is the oft cited measure of "percent proficient." Calculated as a simple percentage of students tested who attained an achievement level of M or E, this subject-specific measure is reported on MDE's website and can be disaggregated from the state level by school, grade, test, and/or student subgroup (MDE, n.d.-a). Though this measure is not directly used by the state to rank schools for recognition or awards, it is frequently used by local media to draw conclusions about the comparative performance of schools and districts, within and across years (Lonetree, 2013; Magan, Webster, & Koumpilova, 2013).

A modified percent proficiency is also the basis for determining AYP (MDE, 2011). In the AYP calculation, state goals are determined based on the percentage of students tested attaining proficiency, where each student having an achievement level of P is counted as an additional half of a proficient student. At the state level, school-level proficiency and achievement gap-reduction goals to be reached by the end of the 2016-

2017 school year are based on the percent of students tested who were proficient in math and reading in 2011 and 2013, respectively. Additionally, the proficiency domain of the MMR is an index based on the success or failure of schools to make AYP subgroup goals (MDE 2012a). As in the case of simple percent proficiency, modified percent proficiency at the school- or state-level does not account for test differences.

2.3.2 Grade-normed percent proficiency. In late 2013, as a part of an attempt to create a list of schools ordered by relative quality to determine the eligibility of existing charter schools to apply for federal Charter Schools Program (CSP) Planning/Implementation grants, a "quality index" based on the three domains of proficiency, growth, and graduation was developed (MDE, n.d.-b). While mirroring MMR in its use of proficiency, the need for a single ranked list of schools, as opposed to the four lists produced in MMR, required the known differences in tests to be accounted for statistically within the proficiency measure. Unfortunately, more sophisticated test-equating methods, either traditional equating methods from classical test theory or item response theory equating, would be inappropriate in this instance as the same constructs are not measured by the tests given in the different grades and subjects (Kolen & Lee, 2011; Zhu, 1998).

A methodology that was different from anything previously reported by MDE was utilized for calculating the comparative proficiency measure in the quality index. Grade-normed percent proficiency or a "proficiency *Z*-score" was calculated for each grade served by each school, with positive grade-level proficiency *Z*-scores indicating above-average proficiency for a specific grade. This linear *Z*-score transformation,

utilizing grade norms, put the locations of proficiency rates for each test in each subject and grade on a single distribution so that the locations could be combined across different tests (Thorndike & Thorndike-Christ, 2010). Within this calculation, proficiency counts for math and reading were combined. Grade-level proficiency *Z*-scores were averaged across all grades served by each school to create a school-level proficiency *Z*-score (MDE, n.d.-b). No indication of research supporting this methodology is provided; however, *Z*-score transformations are commonly used to compare scores from different tests that cannot or should not be equated (Zhu, 1998).

2.3.3 Grade-normed scale scores. Another federal program requiring an annual list of schools ranked by relative school quality is the National Blue Ribbon Schools (NBRS) Program. In the fall of 2013, the NBRS program released a new nomination process asking states to develop their own system of ranking schools that would marry the federal accountability measures defined through their ESEA waivers and the NBRS eligibility requirements (USDOE, n.d.).

Similar to the process for identification of high quality charter schools for federal CSP grant eligibility, NBRS required the creation of a single ordered list of schools, thereby limiting the extent to which MDE could mimic MMR in its methodology. To address the need for scores on all tests across different grades to be comparable, scale scores were normed within grade, test, and subject (K. Edwards, personal communication, January 8, 2014). Thus, these new student-level scale score *Z*-scores had centers at the state student average score for each grade, test, and subject, with positive scores indicating “above average” achievement. Student-level scale score *Z*-scores were

aggregated to the school level within subject. The resulting school-level math and reading scale score *Z*-scores were used, along with other criteria, to rank schools.

This methodology for quantifying school achievement differs from the previous two in two very important ways: (1) the calculation is based on scale scores rather than on achievement levels and (2) the calculation removes all information regarding actually meeting academic standards (proficiency), instead focusing solely on where schools fall in a distribution of scores relative to one another. The first difference is important because it may guard against the potential loss of information incurred from categorizing a continuous variable (MacCallum, Zhang, Preacher, & Rucker, 2002; van Dulmen & Egelund, 2011). The movement of students from D to P or from M to E is not captured by a proficiency calculation but is reflected in scale scores. This means within a proficiency calculation a school is not rewarded for having students increase achievement from D to P or M to E nor is it penalized for having students decrease achievement for E to M or P to D. Masked increases and decreases in student achievement could result in the masking of changes in relative school performance. Therefore, such improvement or decline in student achievement is crucial information to include in the production of a robust relative ranking of schools. The second difference deserves attention because the ultimate goal of ESEA is for all students to reach proficiency on math and reading standards. Ideally, the threshold of proficiency would be captured somehow within a reliable ranking system. Though not currently in use by MDE, a score transformation that grade-norms scale scores with their centers at the cut score for proficiency for each test might

be seen as a hybridization of percent proficiency and scale score transformations. Here, positive scores would indicate "above proficient" achievement.

As in the case of the high quality charter schools, no indication of research supporting the NBRS methodology is provided. However, in a 2008 evaluation of charter schools provided to Minnesota by the Office of the Legislative Auditor (OLA), a nearly identical scale score transformation was used. Citing the different statewide means and standard deviations of the 14 MCA tests administered, scale score data used in the report was standardized within grade and subject to have a mean of 50 and standard deviation of 15 (Nobles et al., 2008).

2.3.4 Large-scale assessments. TIMSS, PISA, and NAEP are three widely known studies in which tests are given nationally (NAEP) and internationally (TIMSS and PISA). Scoring and reporting on each of these tests closely resembles scoring on Minnesota's MCA, MOD, and MTAS exams. Results from TIMSS, PISA, and NAEP tests are also used to create ranked lists of schools according to student performance, though only within test grade and test subject.

2.3.4.1 *TIMSS.* For the TIMSS mathematics and science exams, scale scores with a mean of 500 and a standard deviation of 100 were established at each grade level during the first administration in 1995. Subsequent administrations of the tests were linked to the first administration to allow for longitudinal data analysis. Relative performance of countries within grade level and subject is determined after each administration through a simple mean of scale scores, but relative performance of countries overall is not reported (Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Foy, & Arora, 2012). In addition to

scale score reporting, TIMSS reports at benchmark levels analogous to the four levels of achievement used in Minnesota: Low, Intermediate, High, and Advanced. These benchmarks are determined by scale score cutoffs (Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Foy, & Arora, 2012).

2.3.4.2 *PISA*. PISA exams are administered every three years to 15-year-olds around the globe. Within the content domains of mathematics, reading, and science, scale scores are reported with a mean of 500 and standard deviation of 100. There are also four proficiency levels reported in digital reading, seven proficiency levels reported in paper reading, and six proficiency levels reported in each math and science (OECD, 2013a). Participating countries are ranked on simple average scale scores within subject but, as in TIMSS, no overall ranking across subjects is provided (OECD, 2013b).

2.3.4.3 *NAEP*. The NAEP main assessments differ from both TIMSS and PISA exams in that no individual scale scores are computed. Instead, summary information about subject area scale scores is reported by student groups. Achievement levels of Basic, Proficient, and Advanced are also defined at cut scores within each grade and subject. NAEP also has separate long-term trend assessments in mathematics and reading that have scores reported on a 0-500 scale and in terms of five performance levels set at cut scores of 150, 200, 250, 300, and 350. The long-term trend assessment results are not used to compare states whereas the results of the main assessments are used to examine relative results. Just as in TIMSS and PISA, no ranking across subjects is reported (USDOE, 2014).

2.3.4.4 ***Application to MN assessments and school rankings.*** TIMSS, PISA, and NAEP use results from subject level, grade-specific exams to rank the relative performance of countries and/or states within subjects and grades. Just as with Minnesota test results, scores on TIMSS, PISA and NAEP exams are reported at both scale score and achievement levels. However, none of these results are aggregated across subjects to create a single ranking list of relative performance. Thus, without additional methods in use nationally to consider, it is reasonable to restrict a first investigation into the effects of score transformations on relative rankings to methods, and variations of methods, already in use in Minnesota.

Lack of supporting research leaves it unclear whether there is any advantage in using a specific method of score reporting - proficiency, grade-normed proficiency, grade-normed scale scores, or some other method of score reporting – when calculating an aggregated school-level achievement measure. Using proficiency offers an element of transparency to the public, which is invaluable in highly visible ranking systems such as those used by MDE. Using scale scores prevents the loss of information inherent in the transformation of a continuous measure to a categorical one, potentially resulting in a more accurate and fair ranking system. The actual effects of using these different types of scores still need to be examined.

2.4 Factors Related to Achievement

Using student performance on standardized tests as a proxy for school quality ignores student- and school-level factors, often outside of a school's control, that may be related to student performance on tests. These factors include students' racial/ethnic

backgrounds, students' primary languages, students' family income, and mid-year transfers of students in and out of classrooms and schools. Research exists linking each of these factors to standardized test performance.

2.4.1 Racial/ethnic backgrounds. In Minnesota students are classified as belonging to one of five racial/ethnic groups: American Indian/Alaskan Native (AMI), Asian/Pacific Islander (API), Hispanic (HIS), black and not of Hispanic origin (BLK), and white and not of Hispanic origin (WHT). There is research relating membership in each of the non-white racial/ethnic classifications to decreased expectations in standardized test achievement.

A 2013, analysis of National Assessment of Educational Progress (NAEP) data from 2005, 2007, and 2009, showed students who were identified by the school as American Indian, black, or Hispanic, scored significantly lower than students identified by the school as white on both math and reading tests (Fischer & Stoddard, 2013). In 2011, an analysis of the black-white achievement gap using data from the Educational Longitudinal Study of 2002 showed the most significant predictor of student achievement on standardized tests was socio-economic status (SES), followed closely by race, and that when the sample of students was split by race, SES was an even stronger predictor of black students' test scores (Rowley & Wright, 2011). Similar relationships between SES, ethnicity, and test scores have been found for the Hispanic-white achievement gap (Morales & Saenz, 2007; Madrid, 2011). A 2008 study on the effects of NCLB legislation on American Indian student achievement in Arizona showed that American Indian students performed significantly lower than white students on both math and

reading tests in grades 3, 5, and 8 (Garcia, 2008). Achievement gap literature is often not as concerned with the Asian-white gap (Fischer & Stoddard, 2013); however, proficiency rates by race/ethnicity from the 2013 administration of standardized tests in Minnesota, given in Table 4, show an apparent Asian/Pacific Islander-white gap.

2.4.2 Primary languages. All accountability tests given in Minnesota are given only in English. Not surprisingly, an achievement gap between students who are fluent in English and students who are not fluent in English (often referred to as English Leaners (EL), English Language Learners (ELL), or Limited English Proficient (LEP)) has also been found in research (Martinello, 2008; Turkan & Liu, 2012). Though many researchers suggest this gap may be more of an indication of poor psychometric properties of a test than of actual difference in student knowledge (Young et al., 2008; Abedi, 2004; Abedi, 2002), the relationship between a school's LEP population size and relative ranking may be significant.

Table 4.
Percent proficiency by race/ethnicity

		Percent Proficiency				
Subject	Test name	AMI	API	HIS	BLK	WHT
Reading	MCA-III	35.5	49.7	34.7	33.7	66.0
	MOD-III	35.5	34.1	33.5	31.1	46.1
	MTAS-III	74.6	56.0	69.7	66.7	68.7
Math	MCA-III	39.1	62.1	39.6	36.6	70.7
	MCA-II	28.8	50.6	26.1	23.6	60.5
	MOD-III	8.0	12.4	9.8	7.2	15.3
	MOD-II	16.7	9.2	5.0	4.3	15.1
	MTAS-III	72.4	53.6	66.7	65.5	65.8
	MTAS	65.2	51.4	37.5	48.4	43.6

2.4.3 Income. As the original intent of the ESEA was to close the educational achievement gap between students from low-income families and their peers, it is not surprising that such a gap still exists. Across all fifty states, a negative relationship exists between state average 8th grade NAEP math and reading scores and state child poverty rate (Ladd, 2012). Additional research suggests that the achievement gap between students who qualify for free or reduced price lunch and students who do not widens between kindergarten and grade 12 (Beckman, Messersmith, Shepard, & Cates, 2012). A relationship between race/ethnicity and SES also exists, with higher levels of poverty being found in minority groups, suggesting the effect of SES on standardized test achievement may be best explored as an interaction effect (Strand, 2014).

2.4.4 Student mobility. Mobility is a measure of how often students change schools and/or districts in a year. The relationship of mobility to achievement is more controversial than the factors previously discussed. Evidence exists that students who move schools frequently throughout a school year are likely to see lower achievement than less mobile students (Mehana & Reynolds, 2003; Demie, 2002). However, in Minnesota, test scores for students who are not at a school for the entire school year are not used in school accountability measures (MDE, 2011). Other researchers conclude that students who move classes frequently within a school may also see lower test scores than students who stay in the same classrooms for full terms (Wright, 1999), but this is not a data point Minnesota collects. Further, the effects of mobility on achievement may be confounded with the effects of other factors such as SES and race/ethnicity (Wright, 1999).

From a different perspective on effects of student mobility on achievement, some research has been done examining the effect of a high churn rate in a classroom on the students who are not mobile themselves. Students who have classmates that are tardy are likely to earn lower test scores than those who are in classes with lower rates of tardiness (Gottfried, 2014). Following this line of reasoning, it may be reasonable to expect a school with a higher mobility rate to have lower overall achievement than a school with a lower mobility rate, even though the achievement being measured is that of the non-mobile students.

2.5 Implications for School Ranking

While achievement tests are not the only source of data that should ultimately be aggregated to determine relative school quality, the existence of 38 different standardized tests in Minnesota means achievement test scores may be the most difficult data to transform and aggregate in a fair and robust way. Any method of score transformation and ranking employed needs to balance reliability and accuracy with public transparency and ease of understanding. A lack of research leaves it unclear if it is best to use raw proficiency percentages, grade-normed proficiency percentages, grade-normed scale scores, or some other conversion of the MCA-III, MTAS, and MCA-M scale scores to assess school quality. Ideally, a ranking of schools would also be as independent of demographic characteristics such as race/ethnicity, percent of students in poverty, percent of students receiving English language services, and mobility, as possible.

Federal legislation dating back to the Improving America's Schools Act of 1994 specifies that all students must be held to the same academic standards. This requirement

calls into question whether it is appropriate to create models for ordering schools that control for demographic characteristics. For all students to be treated "equally," all students' test scores should have equal input into the quantification of their school's overall performance. Under these requirements, a "best" school ranking would use the simplest score transformation that was significantly related to the fewest number of school demographics.

2.6 Research Questions

This study will attempt to answer the following three research questions:

1. What effect does the method of transforming student scores (percent proficiency, grade-normed percent proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores) on Minnesota state exams have on relative Minnesota school performance rankings over time?
2. What effect do school demographics (mobility, percent American Indian, percent Asian/Pacific Islander, percent black, percent Hispanic, percent non-white, percent free/reduced price lunch, and percent English learner) have on relative Minnesota school performance rankings over time?
3. What effect do the interactions of method of transforming student scores on Minnesota state exams and school demographics have on relative Minnesota school performance rankings over time?

Chapter 3

Method

3.1 Sample

The data for this study come from Minnesota state accountability tests administered during the 2010-2011, 2011-2012, 2012-2013, and 2013-2014 school years. During each of these years, the Minnesota Comprehensive Assessments (MCAs) were administered in grades 3-8 and 10 in reading and grades 3-8 and 11 in mathematics, the Minnesota Comprehensive Assessments – Modified (MCA-Ms) were administered in grades 5-8 and 10 in reading and grades 5-8 and 11 in mathematics, and the Minnesota Test of Academic Skills (MTAS) was administered in grades 3-8 and 10 in reading and grades 3-8 and 11 in mathematics. Thus, a total of 38 different accountability tests were given each academic year. Additional student demographic data come from state test records and school-level mobility data comes from MDE's website. Following recommendations for maintaining data privacy outlined by the National Center for Education Statistics (NCES) and the public data reporting practices of MDE, only schools from which at least ten academic and demographic data records are available for all four academic years are included in this study (National Center for Education Statistics [NCES], 2010).

Using a complete set of statewide individual level testing and enrollment records affords a unique opportunity for robust analysis. The data consist of approximately 800,000 test records per year coming from approximately 1,900 unique schools, and have as complete demographic information as is available anywhere.

3.2 Procedures and Research Design

This study employed mixed-effects multi-level longitudinal models to study the effects of transformations of student state exam scores and school demographics on relative school ranking. For each school, a relative school ranking was computed for each academic year and used as the outcome in each fitted model. These relative rankings were produced for each of the four score transformation methods under investigation.

3.2.1 Standardized test score transformations. Percent proficiency at the school level was calculated as a ratio of students earning M and E to all students having valid scores on any of the tests in any subject. Grade-normed percent proficiency at a school level was calculated as the grade-size weighted average of grade-normed proficiency, where grade-normed proficiency was the *Z*-score of grade-level percent proficiency by subject. Grade-normed scale scores and proficiency-centered grade-normed scale scores were both averaged to the school level with each student's score(s) given equal weight in the calculations. Completion of these four score transformations resulted in four aggregate measures of school performance for 1,937 schools in 2011, 1,922 schools in 2012, 1,945 schools in 2013, and 1,925 schools in 2014. Once schools with fewer than ten test records in any of the four years, including those having no test records, were removed from the data set, the remaining 1,485 schools, the same schools across all four years, were given ranks based on each of the aggregate measures for each year. Ties in measures of school performance were handled by giving sequential ranks to unique school performance values. Schools having the same aggregate measure of achievement were given the same rank and the school having the next highest aggregate measure of

achievement was given a rank of one greater (worse) than the tied schools. Changes in annual ranks produced using each of the score transformation methods were also calculated for the latter three years by setting the 2011 change in rank to zero for all transformation methods and then finding the difference between each pair of consecutive year's rankings (2014 rank – 2013 rank, 2013 rank – 2012 rank, and 2012 rank – 2011 rank). This method of ranking is similar to the "bracket-rank" method and is used instead of the more common mid-rank method (Kendall, 1945; Student, 1921) for several reasons. First, using mid-rank methods for ties can create non-integer values for ranks. This leads to issues of interpretability when calculating the annual change in rank as a difference between two years of rankings. For example, a school may have had the fourth highest aggregate achievement value in 2011 but have been assigned a mid-rank value of 6.5 based on being tied with five other schools. If, in 2012, the school has the fifth highest aggregate achievement value and does not share that value with any other schools, the change in rank would be a negative movement of 1.5 rank places. A half position change in rank order does not make sense within a measurement aimed at describing strictly ordinal change. The second reason for using sequential ranks instead of mid-ranks has to do with the application of these ranks to the statewide systems of differentiated recognition and support required by ESEA. Even under the current ESEA waivers, states are required to designate certain percentages of ranked schools as either "reward" or "priority." Sequential rankings give smaller rank values (higher rankings) to schools with tied performance measures, effectively prioritizing minimizing false

negative classifications over minimizing false positive classifications. This follows the designation cutpoint procedures used in Minnesota's MMR (MDE, 2014).

The application of the sequential ranking scale to the aggregate measures of school performance appears to transform the interval level measures that were observed to be approximately normally distributed to ordinal level measures, as Stevens (1966) suggested. However, critics of Stevens argue that his identification of only four levels of measurement (nominal, ordinal, interval, and ratio) is over simplistic, and that many of the seemingly ordinal scales used in the social sciences, though not exactly interval, are also not strictly ordinal (Borgatta & Bohrnstedt, 1980). In the application and interpretation of these particular ranks, it is arguable that the distance between each sequential rank is both meaningful and equal in measure. The difference between being ranked first and third is the same as the difference between being ranked thirty-first and thirty-third. As such, even without a continuous measure, the use of parametric statistics can be justified: "[...] because of the robustness of parametric techniques, treating ordinal data as if they were interval would be unlikely to lead to improper conclusions" (Gardner, 1975, p.51). Further, if one can accept the ranks as essentially interval in nature, the measurement of the absolute change in ranking is then a ratio level of measurement as it is both possible and correct to observe a school moving three times as much or half as much as another.

3.2.2 Model covariates. Several demographic variables at the school level were also calculated and used in this study. Percent American Indian (AMI), percent Asian/Pacific Islander (API), percent black (BLK), percent Hispanic (HIS), percent non-

white, percent free/reduced price lunch (FRP), and percent English learner (LEP) were calculated as the number of students across all four years of data with valid test scores having each respective demographic flag within a school divided by the number of students with valid test scores across all four years in that school. It should be noted that these aggregated demographic variables are only as good as the data reported to MDE. In Minnesota, parents are generally asked to identify the racial/ethnic category that best describes each of their children. If a parent does not identify a student's racial/ethnic background, a school employee may identify the student's racial/ethnic background based on the information they have (student name, visual appearance, etc.). To be identified as receiving free/reduced price lunch, a student or parent must fill out an Application for Educational Benefits form from the U.S. Department of Agriculture (USDA) and meet requirements based on household income guidelines. If a student is previously identified as being homeless, a school may choose to identify the student as receiving free/reduced price lunch without completion of the Application for Educational Benefits.

Annual mobility for each school was calculated according to Minnesota Department of Education procedures: the sum of a school's mid-year transfers into a district, mid-year in-district transfers, and mid-year transfers out of a district, divided by the school's K-12 enrollment on October 1st. When schools have small populations and high transfer rates, it is possible to have mobility rates of greater than 100 percent. A weighted averaged of the four years of mobility was calculated to produce a single measure of mobility over time for each school. Of the 1,485 ranked schools, only 1,435

had mobility data available. Distributions of all demographic variables were examined and found to show right skew. Descriptive statistics of all 4-year average school-level demographic variables are given in Table 5.

Notably absent from this list of demographic variables is the percent of testers in a school who receive special education services. The MCA-Ms and MTASs are designed to offer accomodations for such students so the inclusion of this particular demographic variable at the school level would be redundant.

Table 5.
Means, medians, standard deviations, and ranges of average school demographics

Demographic Variable	μ	Median	σ	Minimum	Maximum
Percent non-white	24.9	14.2	26.3	0.0	100.0
Percent AMI	3.4	0.7	11.5	0.0	100.0
Percent API	5.1	1.7	10.2	0.0	99.8
Percent HIS	6.9	3.5	10.3	0.0	98.2
Percent BLK	9.5	2.6	16.6	0.0	100.0
Percent FRP	42.6	38.8	22.8	0.0	100.0
Percent LEP	6.7	1.4	13.6	0.0	100.0
Mobility	17.1	10.0	65.3	0.0	2340.4

3.3 Answering the Research Questions

To answer the first research question – What effect does the method of transforming student scores on Minnesota state exams have on relative Minnesota school performance rankings over time? – preliminary data analysis were conducted and then a multi-level model was fitted (see Equation 1).

3.3.1 Preliminary analysis. Prior to fitting the model, three sets of profile plots were examined. First, plots of the longitudinal rankings produced by each of the four methods of score transformation (percent proficiency, grade-normed percent proficiency,

grade-normed scale scores, and proficiency-centered grade-normed scale scores) were created for a random sample of the schools. These were examined for visible trend differences among score transformation methods and for general shape (linear, quadratic, etc.). Second, plots of the mean four-year school rank for every school produced by each of the four methods were created and examined again for differences among the results produced by each of the four score transformation methods and for possible outlier schools in the data set. Lastly, plots of the standard error of the mean four-year school rank for every school produced by each of the four methods were created and examined for visible differences in magnitude among the score transformation methods. Spearman correlations between rankings were run within year, across score transformation methods and across years, within score transformation method to examine differences between rankings that would not be visibly apparent in the profile plots. Spearman correlations between changes in rankings were also run within year, across score transformation methods to examine differences in the stability of ranking produced using the different score transformations.

Following the procedures used in MMR, the top 15% of schools in the ranked list would be designated "Reward" and the next 25% of school would be designated "Celebration Eligible." Additionally, the bottom 5% of schools would be designated "Priority" and the next lowest 20% of schools would be designated "Continuous Improvement" (MDE, 2014). Ranked lists of schools were compared within year, across score transformation method to determine the percentage overlap of schools in each of these designation categories.

3.3.2 Fitting the first model. A multi-level model with the ordinal ranking produced using each score transformation method at each time as a school-level response variable, time as a level-1 predictor, and the score transformation method as a level-2 categorical predictor variable was fitted to determine if type of score transformation has an effect on school rankings over time. Models utilized an AR(1) (auto regressive) covariance structure because the ranking outputs across all schools at each time point have approximately equal variance and also because under the assumption that test scores are a valid measure of school quality, correlations between two time points should only be dependent on lag. Maximum likelihood estimation was used because of the advantages it holds over other estimation methods in large samples (Raudenbush & Bryk, 2002). As a linear structure was found in the preliminary analysis, this model had the form

$$Y_{ij} = \beta_{0_j} + \beta_{1_j}(time) + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(GrdNrmProf) + \gamma_{02}(GrdNrmScale) + \gamma_{03}(ProfCentScale) + u_{0_j} \quad (1)$$

$$\beta_{1_j} = \gamma_{10} + \gamma_{11}(GrdNrmProf) + \gamma_{12}(GrdNrmScale) + \gamma_{13}(ProfCentScale) + u_{1_j}$$

The significance of each score transformation method was examined to determine if any of the rankings produced by the more complicated score transformation methods (grade-normed percent proficiency, grade-normed scale scores, or proficiency-centered grade-normed scale scores) were significantly different from the rankings produced using percent proficiency. The multi-level model was also compared to a single-level model with time as the only predictor of ranking to assess goodness of fit.

The same multi- and single-level models were fitted to the data using the absolute change in rank as the outcome variable to determine if any of the changes in ranking

produced by the more complicated score transformation methods (grade-normed percent proficiency, grade-normed scale scores, or proficiency-centered grade-normed scale scores) were significantly different from the changes in rankings produced using percent proficiency.

3.3.3 Answering the second research question. To answer the second research question – What effect do school demographics (mobility, percent American Indian, percent Asian/Pacific Islander, percent black, percent Hispanic, percent free/reduced price lunch, and percent English learner) have on relative Minnesota school performance rankings over time? – several multi-level longitudinal regression models were fitted (see Equation 2).

Within each of the four score transformation methods a model using percent non-white (the sum of percent AMI, percent API, percent HIS, and percent BLK), percent free/reduced price lunch (FRP), percent English learner (LEP), and mobility as level-2 predictors and time as a level-1 predictor of ranking was fitted. Each of these four models had the form

$$Y_{ij} = \beta_{0_j} + \beta_{1_j}(time) + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(\%nonWhite) + \gamma_{02}(\%FRP) + \gamma_{03}(\%LEP) + \gamma_{04}(\%mobility) + u_{0_j} \quad (2)$$

$$\beta_{1_j} = \gamma_{10} + \gamma_{11}(\%nonWhite) + \gamma_{12}(\%FRP) + \gamma_{12}(\%LEP) + \gamma_{14}(\%mobility) + u_{1_j}$$

Model coefficients were used to determine what effect, if any, each of these predictors had on school ranking within each type of score transformation. Additionally, models of the same form were fitted using the absolute change in rank produced through each of the four score transformations as the outcome variable. Because the relationships

between percent non-white and rank and percent non-white and absolute change in rank were found to be significant in several of the models, an additional eight models were fitted breaking apart percent non-white into its components of percent AMI, percent API, percent HIS, and percent BLK.

3.3.4 Answering the third research question. To answer the final research question – What effect do the interactions of method of transforming student scores on Minnesota state exams and school demographics have on relative Minnesota school performance rankings over time? – a multi-level model using time as a level-1 predictor, all demographic predictors from the previous analysis and all four score transformation methods as level-2 predictor variables was fitted for both rank and absolute change in rank as outcomes (see Equation 3). To effectively compare differences between transformation methods, models were fitted with each transformation method as the referent group. Significance of level-2 demographic predictors with variance due to score transformation method controlled for was examined and the models were trimmed until only significant predictors remained.

$$Y_{ij} = \beta_{0_j} + \beta_{1_j}(time) + r_{ij}$$

$$\begin{aligned}\beta_{0_j} = {} & \gamma_{00} + \gamma_{01}(trans\ mthd\ 1) + \gamma_{02}(trans\ mthd\ 2) + \gamma_{03}(trans\ mthd\ 3) + \gamma_{04}(demo\ 1) \\ & + \gamma_{05}(demo\ 2) + \gamma_{06}(trans\ mthd\ 1 * demo\ 1) \\ & + \gamma_{07}(trans\ mthd\ 2 * demo\ 1) + \gamma_{08}(trans\ mthd\ 3 * demo\ 1) \\ & + \gamma_{09}(trans\ mthd\ 1 * demo\ 2) + \gamma_{010}(trans\ mthd\ 2 * demo\ 2) \\ & + \gamma_{011}(trans\ mthd\ 3 * demo\ 2) + u_{0_j}\end{aligned} \quad (3)$$

$$\beta_{1_j} = \gamma_{10} + \gamma_{11}(trans\ mthd\ 1) + \gamma_{12}(trans\ mthd\ 2) + \gamma_{13}(trans\ mthd\ 3) + \gamma_{14}(demo\ 1) + \gamma_{15}(demo\ 2) + \gamma_{16}(trans\ mthd\ 1 * demo\ 1) + \gamma_{17}(trans\ mthd\ 2 * demo\ 1) + \gamma_{18}(trans\ mthd\ 3 * demo\ 1) + \gamma_{19}(trans\ mthd\ 1 * demo\ 2) + \gamma_{110}(trans\ mthd\ 2 * demo\ 2) + \gamma_{111}(trans\ mthd\ 3 * demo\ 2) + u_{1_j}$$

Based on the results from each of the three analyses, each score transformation method was evaluated to choose one optimal ranking procedure. In choosing an optimal statewide school ranking system for public consumption it is necessary to consider both how accurate the system is and how easily the methodology behind the ranking system can be understood by the layman. In a system that is accurately measuring relative school quality, it would be reasonable to expect a limited to non-existent relationship between school demographics and school rank. Using proficiency offers some ease of understanding to the public, which is invaluable in highly visible ranking systems such as those used by MDE, while using scale scores prevents the loss of information inherent in the transformation of a continuous measure to a categorical one. Ordered by increasing level of difficulty of understanding, rankings produced by percent proficiency, grade-normed percent proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores were examined for how different each was from rankings produced using each of the other score transformation methods. Each set of rankings was also examined for the strength of its relationship to school demographic characteristics. As an ideal ranking procedure of school quality would strike a balance between a limited relationship with school demographics and simplicity of public understanding of score

transformation method, the ranking system that used the most simplistic score transformation while having the fewest number of significant demographic predictors was considered optimal.

Chapter Four

Results

4.1 Profile Plots

All plots were created using R (R Development Core Team, 2005). Plots of the longitudinal rankings produced by each of the four methods of score transformation (percent proficiency, grade-normed percent proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores) for a random sample of schools indicated that the trends of the rankings under each of the four methods were primarily linear in form (see Figure 1). Plots of the mean four-year school rank for each school produced by each of the four methods showed similar shape and spread across all transformation methods (see Figure 2). Similarly, plots of the standard error of the mean four-year school rank for each school produced by each of the four methods also showed similar shape and spread (see Figure 3).

In summary, all of these profile plots indicated that differences between rankings produced using each of the four methods of score transformation needed to be examined in more depth. High correlations ($r \geq 0.752$ and $r \geq 0.614$, respectively) were found between school rankings within year, across score transformation methods and across years, within score transformation method (see Table 6) as well as between changes in rankings within year, across score transformation methods (see Table 7). Results of the Spearman's rho correlations further supported fitting the first multi-level model to examine differences between rankings and changes in rankings produced by the different methods.

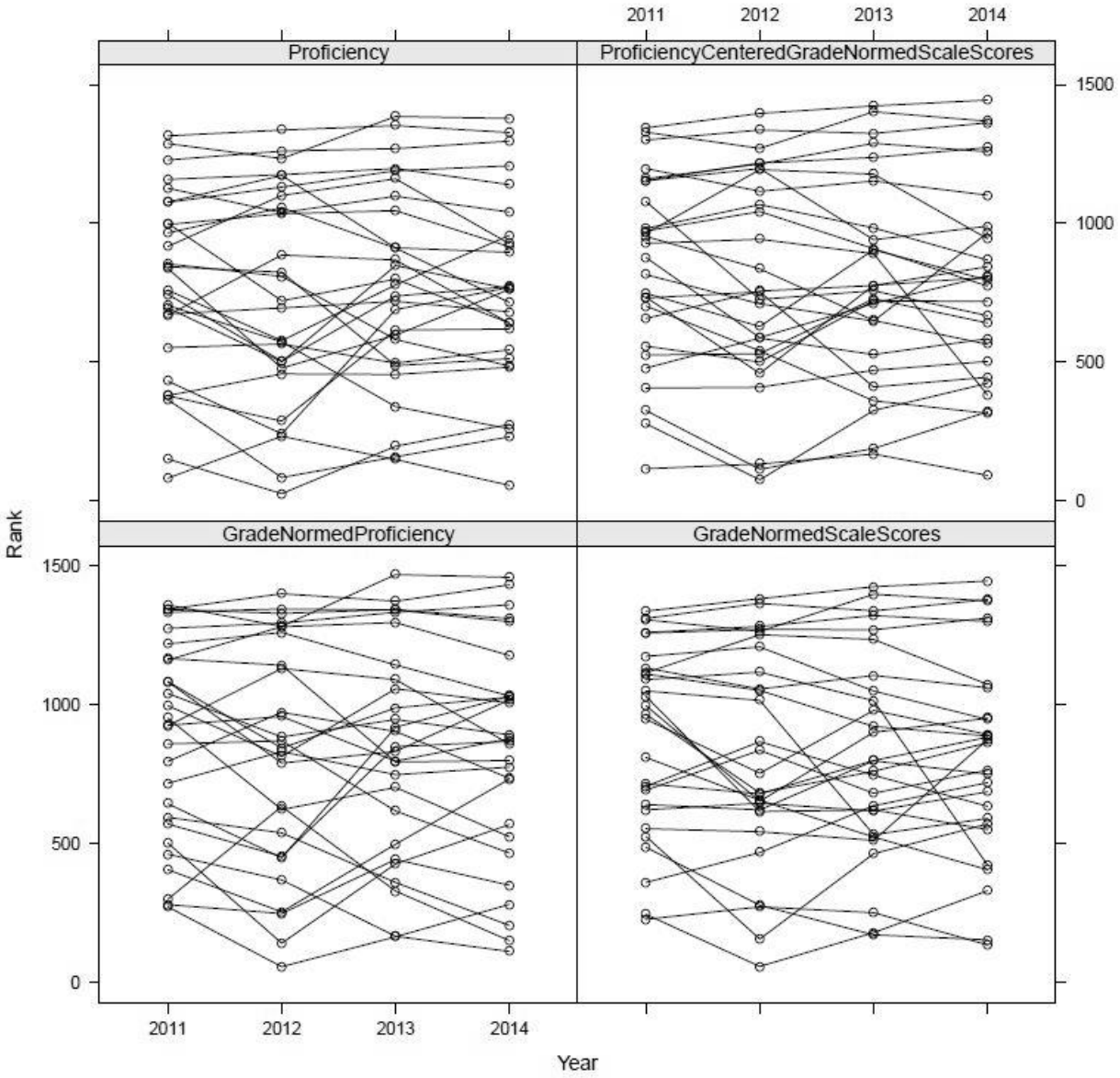

Figure 1. *Rankings produced using each score transformation method over time for a random sample of schools.*

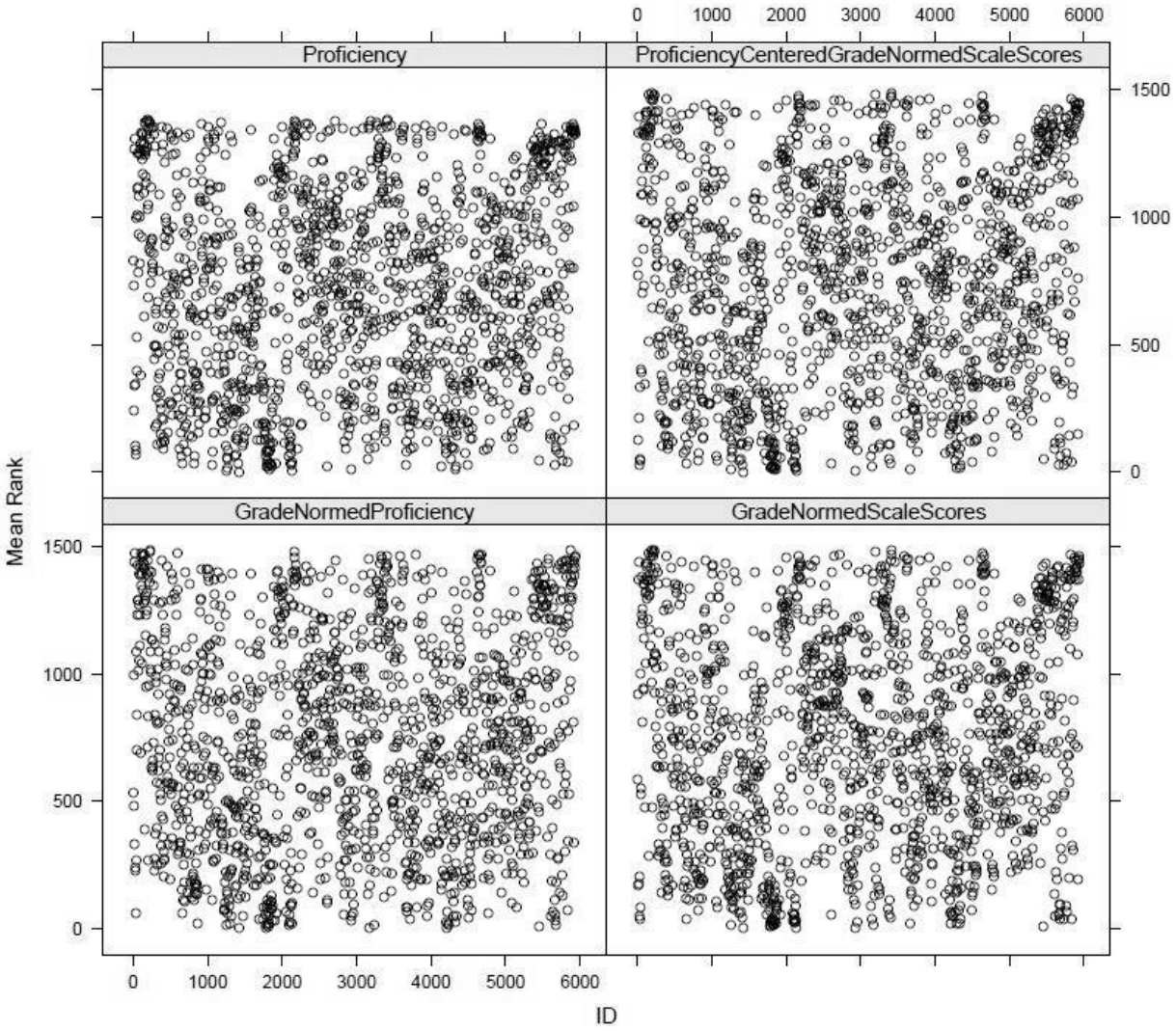

Figure 2. *Mean four-year school rank for each school produced by each of the four score transformation methods.*

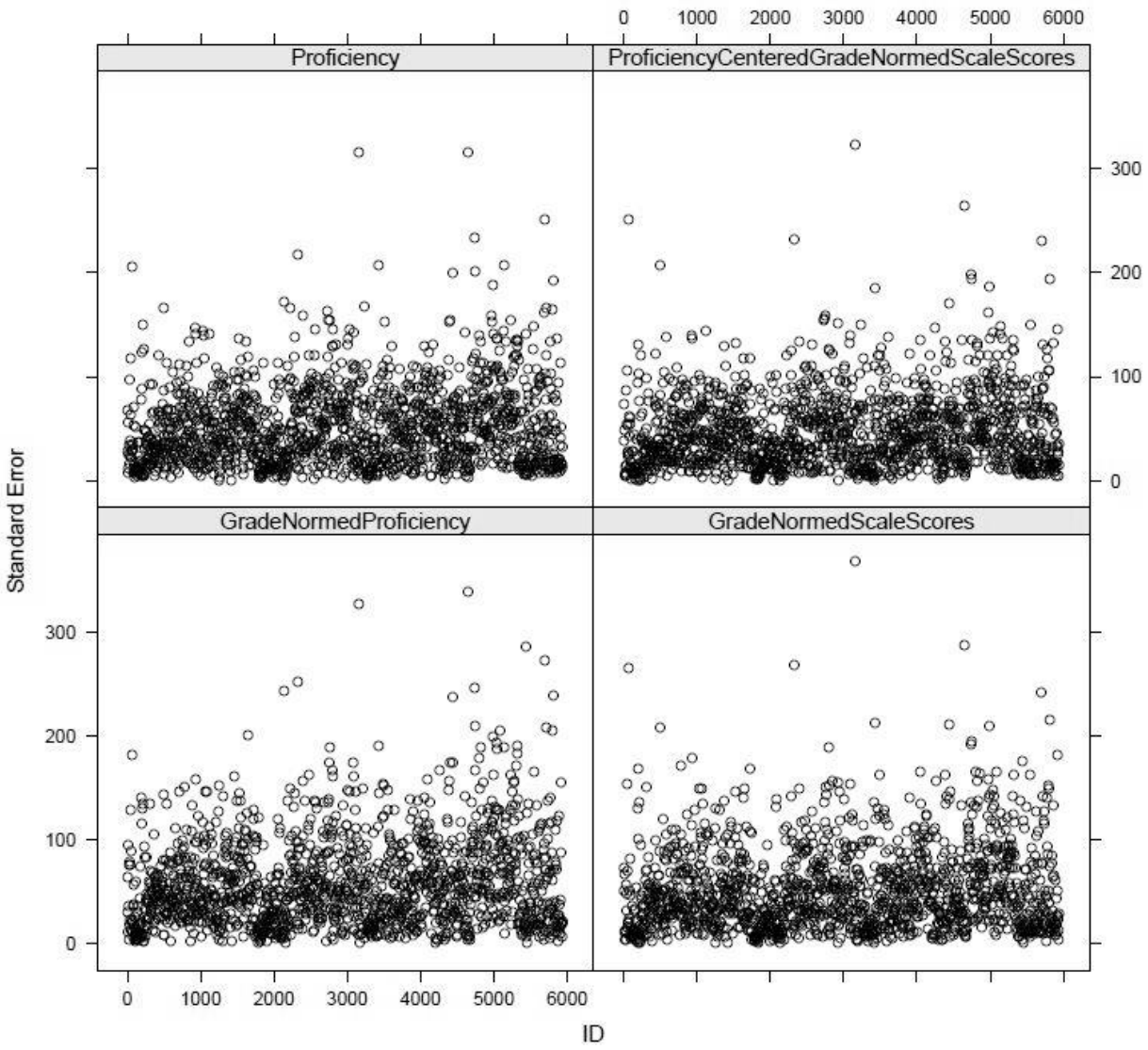

Figure 3. *Standard error of the mean four-year school rank for each school produced by each of the four score transformation methods.*

Table 6.
Spearman's rho correlations between ranks across years and score transformation methods

		2011				2012				2013				2014		
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
2011	1															
	2	.827														
	3	.917	.940													
	4	.985	.790	.922												
2012	1	.914														
	2		.907			.763										
	3			.937		.880	.920									
	4				.953	.976	.752	.916								
2013	1	.898				.888										
	2		.885				.889			.891						
	3			.895				.920		.964	.948					
	4				.911				.924	.989	.878	.974				
2014	1	.873				.877				.934						
	2		.850				.851				.923			.865		
	3			.861				.880				.938		.955	.943	
	4				.890				.907				.946	.988	.847	.946

Note: Score transformation methods are labeled as follows 1 = Percent Proficiency, 2 = Grade-Normed Percent Proficiency, 3 = Grade-Normed Scale Scores, 4 = Proficiency-Centered Grade-=Normed Scale Scores. All correlations were significant at the $p<.01$ level.

Rankings were also examined for overlap of schools in the ranges of each of the MMR designations (Reward, Celebration Eligible, Continuous Improvement, and Priority) across the four score transformation methods within each year. Counts and percentages of overlap are given in Table 8. These overlaps show that the different score transformation methods result in ranks that are different for schools at both ends of the ranking scale.

Table 7.
Spearman's rho correlations between changes in ranks within years across score transformation methods

		2012			2013			2014		
		1	2	3	1	2	3	1	2	3
2012	2	.831								
	3	.668	.731							
	4	.806	.677	.888						
2013	2				.771					
	3				.638	.755				
	4				.844	.614	.788			
2014	2							.939		
	3							.803	.822	
	4							.841	.773	.934

Note: Score transformation methods are labeled as follows 1 = Percent Proficiency, 2 = Grade-Normed Percent Proficiency, 3 = Grade-Normed Scale Scores, 4 = Proficiency-Centered Grade-=Normed Scale Scores. All correlations were significant at the $p<.01$ level.

Table 8.
Total counts and percentages of school overlap in MMR designation categories

Year	Reward (N=223)	Celebration Eligible (N=372)	Continuous Improvement (N=297)	Priority (N=75)
2011	46.6%	19.6%	52.5%	49.3%
2012	41.7%	20.4%	51.5%	41.3%
2013	61.4%	32.3%	58.9%	37.3%
2014	54.3%	29.6%	55.6%	40.0%

4.2 Modeling Differences Amongst Score Transformation Methods

All models were fitted using the HLM 6 software (Raudenbush, Bryk, & Congdon, 2004). Initially, a single-level model,

$$Y_{ij} = \beta_{0_j} + \beta_{1_j}(time) + r_{ij} \ , \quad (4)$$

was fitted to the data. In this model, Y_{ij} is the ranking school *i* received at time *j* = 0, 1, 2, or 3 (2011, 2012, 2013, or 2014 respectively). The school ranking is found by adjusting β_{0_j}, the grand mean of all 23,760 school rankings, by time effects (β_{1_j}) and school residuals (r_{ij}). In this model, each school received four different sets of longitudinal rankings through the four different score transformation methods. Next, a multi-level model of the form

$$Y_{ij} = \beta_{0_j} + \beta_{1_j}(time) + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(GrdNrmProf) + \gamma_{02}(GrdNrmScale) + \gamma_{03}(ProfCentScale) + u_{0_j} \quad (1)$$

$$\beta_{1_j} = \gamma_{10} + \gamma_{11}(GrdNrmProf) + \gamma_{12}(GrdNrmScale) + \gamma_{13}(ProfCentScale) + u_{1_j}$$

was fitted to the data. The same single- and multi-level models were fitted to the data using the absolute value of the change in rank for the latter three years of data as the outcome variable. Parameter and variance component estimates for all four models are given in Table 9. Robust standard errors for fixed effects as well as standard errors for the variance components are also given. Robust standard errors were chosen because of the high number of level-2 units.

Coefficients generated by these models show that there is significantly more spread in the annual rankings produced by grade-normed percent proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores than in those produced using just proficiency rates. Also, while the magnitude of changes in rankings produced using grade-normed proficiency is not significantly different from changes in rankings produced through proficiency, the magnitude of changes in ranks produced by grade-normed scale scores and proficiency-centered grade-normed scale scores are, on

Table 9.
Parameter and variance component estimates with associated robust standard errors, standard errors, and significance for all models

	Rank as Outcome		Absolute Change in Rank as Outcome	
Effect	Single-Level	Multi-Level	Single-Level	Multi-Level
	Fixed effects			
Intercept				
γ_{00}	733.574***	705.249***	122.907***	136.203***
SE	5.470	10.331	2.179	4.479
Time				
γ_{10}	0.605	2.495	-4.302***	-9.0512***
SE	0.963	1.801	0.903	1.770
Grade-Normed Proficiency Intercept				
γ_{01}		37.799*		6.839
SE		15.150		6.752
Grade-Normed Proficiency Slope				
γ_{11}		-2.569		1.546
SE		2.748		2.725
Grade-Normed Scale Score Intercept				
γ_{02}		37.751*		-23.552***
SE		15.199		6.120
Grade-Normed Scale Score Slope				
γ_{12}		-2.495		7.009**
SE		2.699		2.514
Proficiency-Centered Grade-Normed Scale Score Intercept				
γ_{03}		37.751*		-36.473***
SE		15.194		5.675
Proficiency-Centered Grade-Normed Scale Score Slope				
γ_{13}		-2.495		10.442***
SE		2.546		2.351
	Random effects			
Intercept				
$\hat{\tau}_{00}$	168980.651**	168713.211***	5759.981***	5509.712***
SE	*	3257.817	684.502	680.202
	3262.721			
Slope				
$\hat{\tau}_{10}$	3014.466***	3013.274***	81.324	75.927
SE	106.095	106.074	130.848	130.620
Residual				
$\hat{\sigma}^2$	12461.639	12461.639	10002.957	9990.874
SE	161.690	161.690	183.548	183.327
	Fit statistics			
Number of estimated parameters	6	12	6	12
Deviance	319806.228	319797.146	219792.483	219701.948
AIC	319794.228	319773.146	219780.483	219677.948
BIC	319866.683	319918.055	219851.211	219819.405

*$p < .05$, **$p < .01$, ***$p < .001$.

average, significantly less and vary significantly more over time than those produced using proficiency. Though the Akaike Information Criterion (AIC) fit indices indicated that the multi-level models fit better than the single-level models for both rank and absolute change in rank as outcomes, the Bayesian Information Criterion (BIC) indices and deviance statistics suggest that while a multi-level model fits better for absolute change in rank as the outcome, a single-level model fits better for rank as the outcome.

4.3 Modeling Dependence on School Demographics

The relationship between school demographics and relative rankings was examined for each of the four score transformation methods separately. Fifty schools were excluded from the fitting of these models as they had no mobility data available for 2011-2014. For each set of rankings, a single level model (see Equation 4) was fitted using first rank, then absolute change in rank, as the outcome variable. A multi-level model of the form

$$Y_{ij} = \beta_{0_j} + \beta_{1_j}(time) + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(\%nonWhite) + \gamma_{02}(\%FRP) + \gamma_{03}(\%LEP) + \gamma_{04}(\%mobility) + u_{0_j} \quad (2)$$

$$\beta_{1_j} = \gamma_{10} + \gamma_{11}(\%nonWhite) + \gamma_{12}(\%FRP) + \gamma_{12}(\%LEP) + \gamma_{14}(\%mobility) + u_{1_j}$$

was also fitted for each of the outcome variables produced through each of the score transformation methods. Parameter and variance component estimates, robust standard errors for fixed effects, and standard errors for the variance components for all four models of the outcomes generated using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores are given in Tables 10-13, respectively.

Coefficients in these models show that school ranks produced using grade-normed proficiency are dependent on the fewest number of demographic predictors: only on percent FRP. Rankings produced using proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores are also all dependent on the percent of students flagged as FRP. Additionally, rankings produced using proficiency and proficiency-centered grade-normed scale scores are dependent on the percent of students flagged as LEP and rankings produced using grade-normed scale scores are dependent on the percent of students who were identified as non-white. In each case, fit statistics indicated the multi-level model fit the outcome variable better than the corresponding single-level model.

Absolute changes in rankings over time produced using each of the score transformation methods were significantly related to the percent of students who were identified as AMI, API, HIS or BLK. Additionally, the absolute changes in ranking produced using grade-normed scale scores were significantly related to the percent of students flagged as FRP, and the absolute changes in rankings produced using grade-normed proficiency were significantly related to the percent of students flagged as LEP. Change in the magnitude of changes in rankings over time were significantly related to percent mobility for rankings produced using proficiency, grade-normed proficiency, and grade-normed scale scores; to the percent of students identified as FRP for rankings produced using proficiency and grade-normed scale scores; to the percent of students identified as non-white for rankings produced using grade-normed scale scores; and to

Table 10.
Parameter and variance component estimates with associated robust standard errors, standard errors, and significance for models of rankings and absolute change in rankings produced using proficiency

	Rank as Outcome		Absolute Change in Rank as Outcome	
Effect	Single-Level	Multi-Level	Single-Level	Multi-Level
		Fixed effects		
Intercept				
γ_{00}	684.481***	201.393***	138.622***	175.731***
SE	10.256	16.331	4.520	9.363
Time				
γ_{10}	2.735	-8.279*	-9.405***	-20.084***
SE	1.837	3.389	1.769	3.360
% NonWHT Intercept				
γ_{01}		0.157		-1.043***
SE		0.480		0.256
% NonWHT Slope				
γ_{11}		-0.013		-0.131
SE		0.117		0.102
% FRP Intercept				
γ_{02}		11.512***		-0.321
SE		0.540		0.286
% FRP Slope				
γ_{12}		0.278*		0.368**
SE		0.118		0.107
% LEP Intercept				
γ_{03}		-1.602*		0.399
SE		0.642		0.331
% LEP Slope				
γ_{13}		-0.006		-0.181
SE		0.166		0.120
% Mobility Intercept				
γ_{04}		0.434		-0.044
SE		0.318		0.046
% Mobility Slope				
γ_{14}		-0.016		-0.018**
SE		0.010		0.005
		Random effects		
Intercept				
$\hat{\tau}_{00}$	141616.598***	75514.677***	8485.963***	7793.021***
SE	5640.973	3177.401	1435.045	1407.026
Slope				
$\hat{\tau}_{10}$	2173.021***	2138.440***	137.105	157.711
SE	193.982	192.780	259.381	258.379
Residual				
$\hat{\sigma}^2$	13341.710	13341.710	9687.67	9628.780
SE	352.197	352.197	361.667	359.468
		Fit statistics		
Number of estimated parameters	6	14	6	14
Deviance	77057.253	76072.171	53092.124	52928.133
AIC	77045.253	76044.171	53080.124	52900.133
BIC	77109.184	76193.344	53142.331	53045.282

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 11.

Parameter and variance component estimates with associated robust standard errors, standard errors, and significance for models of rankings and absolute change in rankings produced using grade-normed proficiency

	Rank as Outcome		Absolute Change in Rank as Outcome	
Effect	Single-Level	Multi-Level	Single-Level	Multi-Level
		Fixed effects		
Intercept				
γ_{00}	721.062***	129.184***	145.275***	170.848***
SE	11.001	16.126	5.068	9.854
Time				
γ_{10}	0.339	0.315	-7.940***	-16.451***
SE	2.117	3.954	2.066	3.899
% NonWHT Intercept				
γ_{01}		-0.288		-1.730***
SE		0.461		0.282
% NonWHT Slope				
γ_{11}		0.020		0.054
SE		0.129		0.119
% FRP Intercept				
γ_{02}		14.089***		0.308
SE		0.528		0.315
% FRP Slope				
γ_{12}		0.013		0.241
SE		0.132		0.123
% LEP Intercept				
γ_{03}		0.401		0.610*
SE		0.674		0.305
% LEP Slope				
γ_{13}		-0.090		-0.364**
SE		0.159		0.138
% Mobility Intercept				
γ_{04}		0.323		-0.008
SE		0.238		0.034
% Mobility Slope				
γ_{14}		-0.023		-0.026**
SE		0.013		0.007
		Random effects		
Intercept				
$\hat{\tau}_{00}$	162544.383***	61001.283***	9009.757***	7936.910***
SE	6489.776	2708.205	1771.447	1733.707
Slope				
$\hat{\tau}_{10}$	3255.665***	3252.815***	211.622	230.558
SE	254.306	254.205	330.826	330.031
Residual				
$\hat{\sigma}^2$	15877.495	15877.495	12318.754	12269.349
SE	419.137	419.137	459.892	458.048
		Fit statistics		
Number of estimated parameters	6	14	6	14
Deviance	78122.595	76676.111	53954.779	53759.475
AIC	78110.595	76648.111	53942.779	53731.475
BIC	78174.526	76797.284	54004.986	53876.624

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 12.

Parameter and variance component estimates with associated robust standard errors, standard errors, and significance for models of rankings and absolute change in rankings produced using grade-normed scale scores

	Rank as Outcome		Absolute Change in Rank as Outcome	
Effect	Single-Level	Multi-Level	Single-Level	Multi-Level
		Fixed effects		
Intercept				
γ_{00}	720.580***	112.740***	115.082***	123.685***
SE	11.069	16.052	4.253	8.073
Time				
γ_{10}	0.337	4.199	-2.442	-6.801*
SE	2.062	3.691	1.802	3.311
% NonWHT Intercept				
γ_{01}		-1.642**		-1.472***
SE		0.476		0.240
% NonWHT Slope				
γ_{11}		0.262*		-0.037
SE		0.129		0.104
% FRP Intercept				
γ_{02}		15.377***		0.648*
SE		0.546		0.260
% FRP Slope				
γ_{12}		-0.225		0.144
SE		0.128		0.107
% LEP Intercept				
γ_{03}		-0.396		0.170
SE		0.611		0.255
% LEP Slope				
γ_{13}		-0.132		-0.085
SE		0.153		0.104
% Mobility Intercept				
γ_{04}		0.359		-0.047
SE		0.275		0.042
% Mobility Slope				
γ_{14}		0.001		-0.011
SE		0.019		0.007
		Random effects		
Intercept				
$\hat{\tau}_{00}$	167856.866***	66380.924***	4486.845***	3688.561***
SE	6564.583	2780.727	1292.710	1269.176
Slope				
$\hat{\tau}_{10}$	3843.004***	3824.965***	85.450	85.792
SE	235.335	234.684	252.729	252.574
Residual				
$\hat{\sigma}^2$	11278.100	11278.100	9487.895	9481.693
SE	297.721	297.721	354.208	353.977
		Fit statistics		
Number of estimated parameters	6	14	6	14
Deviance	77088.252	75702.741	52790.635	52594.342
AIC	77076.252	75694.741	52778.635	52566.342
BIC	77140.183	75823.914	52840.842	52711.491

*p < .05, **p < .01, ***p < .001.

Table 13.
Parameter and variance component estimates with associated robust standard errors, standard errors, and significance for models of rankings and absolute change in rankings produced using proficiency-centered grade-normed scale scores

	Rank as Outcome		Absolute Change in Rank as Outcome	
Effect	Single-Level	Multi-Level	Single-Level	Multi-Level
		Fixed effects		
Intercept				
γ_{00}	720.395***	184.050***	102.045***	113.107***
SE	11.056	17.574	3.543	6.707
Time				
γ_{10}	0.275	-8.148*	1.079	-1.805
SE	1.846	3.334	1.557	2.811
% NonWHT Intercept				
γ_{01}		-0.855		-0.832***
SE		0.513		0.214
% NonWHT Slope				
γ_{11}		0.139		-0.187*
SE		0.119		0.092
% FRP Intercept				
γ_{02}		13.382***		0.238
SE		0.582		0.219
% FRP Slope				
γ_{12}		0.120		0.185*
SE		0.117		0.092
% LEP Intercept				
γ_{03}		-1.698*		0.066
SE		0.657		0.258
% LEP Slope				
γ_{13}		0.030		0.004
SE		0.161		0.101
% Mobility Intercept				
γ_{04}		0.454		-0.019
SE		0.337		0.037
% Mobility Slope				
γ_{14}		-0.013		-0.019***
SE		0.009		0.004
		Random effects		
Intercept				
$\hat{\tau}_{00}$	168644.429***	89779.932***	1902.29	1573.008
SE	6550.537	3608.296	1021.206	1011.596
Slope				
$\hat{\tau}_{10}$	2958.818***	2921.526***	63.645	54.563
SE	189.484	188.144	209.593	209.242
Residual				
$\hat{\sigma}^2$	9649.269	9649.269	7906.432	7902.824
SE	254.723	254.723	294.661	294.516
		Fit statistics		
Number of estimated parameters	6	14	6	14
Deviance	76341.881	75309.562	52016.000	51861.821
AIC	76329.881	75281.562	52004.000	51833.821
BIC	76393.812	75430.735	52066.207	51978.970

*$p < .05$, **$p < .01$, ***$p < .001$.

the percent of students identified as LEP for rankings produced using grade-normed proficiency.

The significant relationship of the percent of students who were identified as belonging to any ethnic or racial group other than white showed that further investigation was warranted. Multi-level models were fitted to rankings and absolute changes in rankings produced by all score transformation methods. Parameter and variance component estimates, along with associated robust standard errors and significance are reported in Tables 14-17.

Again, rankings produced using each of the score transformation methods were significantly related to the percent of students at a school who were flagged as FRP. Rankings produced using proficiency and grade-normed proficiency were significantly related to the percent of students identified as HIS, rankings produced using proficiency and proficiency-centered grade-normed scale scores were significantly related to the percent of students flagged as LEP, and rankings produced using grade-normed scales scores were significantly related to the percent of students identified as either API or BLK.

Absolute changes in rankings produced by all score transformation methods were significantly related to the percent of students identified as API, HIS, or BLK. Absolute changes in rankings produced by grade-normed proficiency and grade-normed scale scores were also significantly related to the percent of students identified as AMI. Absolute changes in rankings produced using proficiency were significantly related to the

Table 14.

Parameter estimates with associated robust standard errors and significance for models of rankings produced using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores

Effect	Proficiency	Grade-normed proficiency	Grade-normed scale scores	Proficiency-centered grade-normed scale scores
		Fixed effects		
Intercept				
γ_{00}	192.604***	124.814***	107.174***	176.892***
SE	18.260	18.586	18.052	19.471
Time				
γ_{10}	-7.209	0.952	4.582	-7.262*
SE	3.683	4.353	3.989	3.572
% AMI Intercept				
γ_{01}	-0.084	0.119	-1.189	-0.916
SE	0.669	0.580	0.620	0.703
% AMI Slope				
γ_{11}	0.152	0.127	0.286	0.254
SE	0.177	0.176	0.176	0.188
% API Intercept				
γ_{02}	-0.011	-0.947	-2.201**	-1.163
SE	0.803	0.797	0.752	0.845
% API Slope				
γ_{12}	-0.121	-0.082	0.163	0.007
SE	0.179	0.195	0.163	0.160
% HIS Intercept				
γ_{03}	2.991**	2.432*	1.284	1.830
SE	0.911	0.954	0.907	0.952
% HIS Slope				
γ_{13}	-0.056	0.066	0.336	0.168
SE	0.218	0.232	0.215	0.209
% BLK Intercept				
γ_{04}	-0.044	-0.614	-2.052***	-1.075
SE	0.535	0.523	0.538	0.573
% BLK Slope				
γ_{14}	-0.049	-0.002	0.276	0.126
SE	0.142	0.153	0.160	0.148

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 14 Cont.

Effect	Proficiency	Grade-normed proficiency	Grade-normed scale scores	Proficiency-centered grade-normed scale scores
		Fixed effects		
% FRP Intercept				
γ_{05}	11.516***	13.967***	15.267***	13.351***
SE	0.573	0.571	0.580	0.614
% FRP Slope				
γ_{15}	0.248*	-0.011	-0.271	0.091
SE	0.125	0.141	0.136	0.124
% LEP Intercept				
γ_{06}	-2.682**	-0.194	-1.084	-2.600**
SE	0.975	1.066	0.946	0.986
% LEP Slope				
γ_{16}	0.116	-0.027	-0.114	0.114
SE	0.225	0.233	0.200	0.195
% Mobility Intercept	0.428	0.320	0.356	0.449
γ_{07}	0.314	0.235	0.271	0.333
SE				
% Mobility Slope	-0.015	-0.023	0.000	-0.012
γ_{17}	0.010	0.013	0.018	0.009
SE				

*p < .05, **p < .01, ***p < .001.

Table 15.

Variance component estimates with associated standard errors and significance for models of rankings produced using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores

Effect	Proficiency	Grade-normed proficiency	Grade-normed scale scores	Proficiency-centered grade-normed scale scores
		Random effects		
Intercept				
$\hat{\tau}_{00}$	74875.239***	60296.378***	65582.054***	89185.671***
SE	3153.601	2682.045	2750.988	3586.138
Slope				
$\hat{\tau}_{10}$	2134.559***	3250.451***	3823.486***	2918.687***
SE	192.645	254.122	234.631	188.041
Residual				
$\hat{\sigma}^2$	13341.710	15877.495	11278.100	9649.269
SE	352.197	419.137	297.721	254.723
		Fit statistics		
Number of estimated parameters	20	20	20	20
Deviance	76058.301	76655.997	75681.909	75296.871
AIC	76018.301	76615.997	75641.909	75256.871
BIC	76231.405	76829.101	75855.013	75469.975

*p < .05, **p < .01, ***p < .001.

percent of students flagged as LEP, and absolute changes in rankings produced by grade-normed scale scores were significantly related to the percent of students flagged as FRP.

AIC and BIC indices were used to compare each of the models that used percent AMI, API, HIS, and BLK as predictors to the models that used percent non-white as predictors. For rankings and absolute changes in rankings produced using proficiency and grade-normed proficiency, the AIC favored the more complex models while the BIC favored the simpler models. For rankings and absolute changes in rankings produced using grade-normed scales scores both the AIC and BIC indicated the more complex model was a better fit, and for rankings and absolute changes in rankings produced using produced using proficiency-centered grade-normed scale scores both the AIC and BIC indicated the simpler model was a better fit. Partly due to these fit indices and partly due to the increased relationship between ranks and racial/ethnic covariates that comes from splitting school populations out by specific racial/ethnic categories, a decision was made to return to using percent non-white as a covariate in the remaining analyses.

4.4 Modeling Interactions of School Demographics and Score Transformation Methods

To examine the interaction effects of score transformation method and demographic characteristics, four models were fitted using the demographic predictors percent non-white, percent FRP, percent LEP, mobility, and the interactions of each of these demographic predictors with three score transformations methods to predict rank over time. Percent non-white was used in lieu of the separate measures of percent AMI, percent API, percent HIS, and percent BLK because, according to the BIC, the simpler

Table 16.
Parameter estimates with associated robust standard errors and significance for models of absolute change in rankings produced using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores

Effect	Proficiency	Grade-normed proficiency	Grade-normed scale scores	Proficiency-centered grade-normed scale scores
		Fixed effects		
Intercept				
γ_{00}	183.969***	173.322***	126.613***	118.258***
SE	9.869	10.512	8.700	7.079
Time				
γ_{10}	-22.169***	-17.653***	-7.586*	-3.029
SE	3.578	4.315	3.513	2.958
% AMI Intercept				
γ_{01}	-0.630	*-1.833***	-1.281***	-0.341
SE	0.426	*0.397*	0.350	0.392
% AMI Slope				
γ_{11}	-0.167	0.090	-0.097	-0.314*
SE	0.150	0.155	0.132	0.136
% API Intercept				
γ_{02}	-2.113***	-2.096***	-1.955***	-1.646***
SE	0.338	0.360	0.302	0.261
% API Slope				
γ_{12}	0.161	0.252	0.095	0.003
SE	0.125	0.157	0.127	0.136
% HIS Intercept				
γ_{03}	-1.407**	-1.881***	-1.396**	-0.743*
SE	0.425	0.448	0.401	0.345
% HIS Slope				
γ_{13}	-0.040	0.083	-0.068	-0.222
SE	0.166	0.202	0.160	0.143
% BLK Intercept				
γ_{04}	-0.828**	-1.537***	-1.409***	-0.858***
SE	0.277	0.314	0.267	0.224
% BLK Slope				
γ_{14}	-0.226	-0.035	-0.050	-0.186
SE	0.116	0.139	0.122	0.106

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 16 cont.

Effect	Proficiency	Grade-normed proficiency	Grade-normed scale scores	Proficiency-centered grade-normed scale scores
		Fixed effects		
% FRP Intercept				
γ_{05}	-0.516	0.261	0.561*	0.078
SE	0.298	0.330	0.273	0.228
% FRP Slope				
γ_{15}	0.415***	0.268*	0.169	0.224*
SE	0.112	0.131	0.112	0.096
% LEP Intercept				
γ_{06}	1.091**	0.756	0.395	0.520
SE	0.400	0.400	0.347	0.303
% LEP Slope				
γ_{16}	-0.335*	-0.432*	-0.147	-0.104
SE	0.146	0.197	0.135	0.116
% Mobility Intercept				
γ_{07}	-0.043	-0.010	-0.047	-0.018
SE	0.046	0.035	0.042	0.037
% Mobility Slope				
γ_{17}	-0.017**	-0.026**	-0.011	-0.020***
SE	0.005	0.007	0.007	0.004

*p < .05, **p < .01, ***p < .001.

Table 17.

Variance component estimates with associated standard errors and significance for models of absolute change in rankings produced using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores

Effect	Proficiency	Grade-normed proficiency	Grade-normed scale scores	Proficiency-centered grade-normed scale scores
		Random effects		
Intercept				
$\hat{\tau}_{00}$	7711.108***	7933.770***	3671.817***	1532.652
SE	1404.018	1733.132	1268.541	1009.622
Slope				
$\hat{\tau}_{10}$	154.553	229.834	85.252	56.571
SE	258.174	329.907	252.526	209.111
Residual				
$\hat{\sigma}^2$	9624.232	12265.406	9480.387	7895.917
SE	359.298	457.900	353.928	294.256
		Fit statistics		
Number of estimated parameters	20	20	20	20
Deviance	52921.068	53758.234	52592.794	51854.332
AIC	52881.068	53718.234	52552.794	51814.332
BIC	53088.419	53925.585	52760.145	52021.683

*p < .05, **p < .01, ***p < .001.

models built in the analysis for the second research question had better fit for three out of four score transformation methods. To effectively compare differences in main and interaction effects between rankings generated using different score transformations, the different score transformation methods were each used as the referent group in a model.

In the first four full models, no significant relationship was found between either mobility or the interaction of mobility with any of the score transformation methods and school rank over time. Four trimmed models were then fitted with mobility removed. This time percent LEP and the interactions between LEP and each of the score transformation methods were found to be non-significant predictors in each model. Four additional trimmed models were then fitted with percent LEP removed. The parameter estimates, along with associated robust standard errors and significance from these four models are reported in Table 18. Distributions of the level-1 and level-2 residuals for the final model were examined. Level-1 residuals were found to be reasonably normal distributed with mean 0 and variance of $\hat{\sigma}^2$. Level-2 residuals exhibited a uniform distribution with mean 0 and variance $\hat{\tau}_{00}$.

The rankings produced by both proficiency and grade-normed proficiency showed the least dependence on demographic factors as each was only dependent on the percent of students flagged as FRP. Rankings produced using grade-normed scale scores and proficiency-centered grade-normed scale scores were additionally dependent on the percent of students identified as non-white. Rankings produced through grade-normed proficiency were significantly more dependent on the percent of students identified as FRP than those produced through proficiency.

Table 18.
Parameter estimates with associated robust standard errors and significance for models of rankings produced using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores

Effect	Proficiency (Method 1)	Grade-normed proficiency (Method 2)	Grade-normed scale scores (Method 3)	Proficiency-centered grade-normed scale scores (Method 4)
		Fixed effects		
Intercept				
γ_{00}	215.660***	141.089***	128.852***	200.523***
SE	16.590	16.573	16.563	17.899
Time				
γ_{10}	-7.760*	0.611	4.097	-8.294**
SE	3.239	3.773	3.499	3.153
% non WHT Intercept				
γ_{01}	-0.472	-0.173	-1.717***	-1.454**
SE	0.423	0.406	0.423	0.454
% non WHT Slope				
γ_{11}	0.014	0.016	0.212*	0.158
SE	0.099	0.110	0.108	0.099
% FRP Intercept				
γ_{02}	11.775***	14.239***	15.428***	13.591***
SE	0.515	0.511	0.524	0.554
% FRP Slope				
γ_{12}	0.233*	-0.025	-0.220	0.103
SE	0.111	0.124	0.119	0.108
Method 1 Intercept				
γ_{0j}	---	74.570**	86.808***	15.136
SE	---	23.450	23.442	24.405
Method 1 Slope				
γ_{1j}	---	-8.371	-11.857*	0.534
SE	---	4.973	4.768	4.520
Method 2 Intercept				
γ_{0k}	-74.570**	---	12.238	-59.434*
SE	23.450	---	23.431	24.394
Method 2 Slope				
γ_{1k}	8.371	---	-3.486	8.905
SE	4.973	---	5.146	4.917
Method 3 Intercept				
γ_{0m}	-86.808***	-12.238	---	-71.671**
SE	23.442	23.431	---	24.386
Method 3 Slope				
γ_{1m}	11.857*	3.486	---	12.391**
SE	4.768	5.146	---	4.710
Method 4 Intercept				
γ_{0n}	-15.136	59.434*	71.671**	---
SE	24.405	24.394	24.386	---
Method 4 Slope				
γ_{1n}	-0.534	-8.905	-12.391**	---
SE	4.520	4.917	4.710	---

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 18 cont.

Effect	Proficiency (Method 1)	Grade-normed proficiency (Method 2)	Grade-normed scale scores (Method 3)	Proficiency-centered grade-normed scale scores (Method 4)
		Fixed effects		
nWHT x Mthd1Intercept				
γ_{0p}	---	-0.299	1.245*	0.982
SE	---	0.586	0.598	0.620
nWHT x Mthd1 Slope				
γ_{1p}	---	-0.002	-0.199	-0.144
SE	---	0.148	0.146	0.140
nWHT x Mthd2 Intercept				
γ_{0q}	0.299	---	1.545**	1.281*
SE	0.586	---	0.586	0.609
nWHT x Mthd2 Slope				
γ_{1q}	0.002	---	-0.197	-0.142
SE	0.148	---	0.154	0.148
nWHT x Mthd3 Intercept				
γ_{0r}	-1.245*	-1.545**	---	-0.264
SE	0.598	0.586	---	0.621
nWHT x Mthd3 Slope				
γ_{1r}	0.199	0.197	---	0.055
SE	0.146	0.154	---	0.146
nWHT x Mthd4 Intercept				
γ_{0s}	-0.982	-1.281*	0.264	---
SE	0.620	0.609	0.621	---
nWHT x Mthd4 Slope				
γ_{1s}	0.144	0.142	-0.055	---
SE	0.140	0.148	0.146	---
FRP x Mthd1 Intercept				
γ_{0t}	---	-2.464**	-3.653***	-1.816*
SE	---	0.725	0.734	0.756
FRP x Mthd1 Slope				
γ_{1t}	---	0.258	0.453**	0.130
SE	---	0.166	0.162	0.155
FRP x Mthd2 Intercept				
γ_{0u}	2.464**	---	-1.189	0.648
SE	0.725	---	0.731	0.754
FRP x Mthd2 Slope				
γ_{1u}	-0.258	---	0.195	-0.128
SE	0.166	---	0.172	0.165
FRP x Mthd3 Intercept				
γ_{0v}	3.653***	1.189	---	1.837
SE	0.734	0.731	---	0.762
FRP x Mthd3 Slope				
γ_{1v}	-0.453**	-0.195	---	-0.323*
SE	0.162	0.172	---	0.161
FRP x Mthd4 Intercept				
γ_{0w}	1.816*	-0.648	-1.837*	---
SE	0.756	0.754	0.762	---
FRP x Mthd4 Slope				
γ_{1w}	-0.130	0.128	0.323*	---
SE	0.155	0.165	0.161	---

*$p < .05$, **$p < .01$, ***$p < .001$.

Four models using all demographic predictors, three score transformations, and the interactions of score transformations and demographic characteristics were also built for absolute change in rank as the outcome variable. In all models, mobility, the interactions of mobility and score transformation methods, and the interactions of percent LEP and score transformation methods were not significant predictors of intercept. The interactions of percent non-white and score transformation methods, percent FRP and score transformation methods, and mobility and score transformation methods were also not found to be significant predictors of change in absolute change in rank over time (slope). Predictors that were not significant in any model were removed and the reduced models were fitted. After the second set of models were fit, the interaction of percent LEP and score transformation methods was also not a significant predictor of the slope of the absolute change in rank over time and was removed from the models. Parameter estimates, along with associated robust standard errors and significance from the four reduced models are reported in Table 19.

Estimated coefficients from these models indicate that absolute changes in rankings produced using proficiency are generally greater and show less variation over time than those produced using either grade-normed scale scores or proficiency-centered grade-normed scale scores. Absolute changes in rankings produced using grade-normed proficiency and grade- normed scale scores are significantly more dependent on the percent of students identified as non-white and the percent of students identified as FRP respectively, than absolute changes in rankings produced using proficiency.

Table 19.
Parameter estimates with associated robust standard errors and significance for models of absolute changes in rankings produced using proficiency, grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores

Effect	Proficiency (Method 1)	Grade-normed proficiency (Method 2)	Grade-normed scale scores (Method 3)	Proficiency-centered grade-normed scale scores (Method 4)
		Fixed effects		
Intercept				
γ_{00}	166.758***	166.810***	127.807***	119.972***
SE	7.221	7.562	6.352	5.655
Time				
γ_{10}	-15.749***	-14.285***	-8.786***	-5.265*
SE	2.264	2.508	2.298	2.076
% non WHT Intercept				
γ_{01}	-1.300***	-1.665***	-1.548***	-1.218***
SE	0.121	0.128	0.114	0.107
% FRP Intercept				
γ_{02}	0.016	0.387*	0.528**	0.211
SE	0.174	0.183	0.157	0.147
% FRP Slope				
γ_{12}	0.200***	0.200***	0.200***	0.200***
SE	0.039	0.039	0.039	0.039
% LEP Intercept				
γ_{03}	0.461**	0.461**	0.461**	0.461**
SE	0.144	0.144	0.144	0.144
% LEP Slope				
γ_{13}	-0.228***	-0.228***	-0.228***	-0.228***
SE	0.056	0.056	0.056	0.056
% Mobility Slope				
γ_{14}	-0.030***	-0.030***	-0.030***	-0.030***
SE	0.007	0.007	0.007	0.007
Method 1 Intercept				
γ_{0j}	---	-0.053	38.950***	46.786***
SE	---	9.437	8.534	8.125
Method 1 Slope				
γ_{1j}	---	-1.465	-6.963**	-10.484***
SE	---	2.713	2.521	2.353
Method 2 Intercept				
γ_{0k}	0.053	---	39.003***	46.839***
SE	9.437	---	8.931	8.541
Method 2 Slope				
γ_{1k}	1.465	---	-5.498*	-9.019**
SE	2.713	---	2.738	2.583

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 19 cont.

Effect	Proficiency (Method 1)	Grade-normed proficiency (Method 2)	Grade-normed scale scores (Method 3)	Proficiency-centered grade-normed scale scores (Method 4)
		Fixed effects		
Method 3 Intercept				
γ_{0m}	-38.950***	-39.003***	---	7.836
SE	8.534	8.931	---	7.531
Method 3 Slope				
γ_{1m}	6.963**	5.498*	---	-3.521
SE	2.521	2.738	---	2.381
Method 4 Intercept				
γ_{0n}	-46.786***	-46.839***	-7.836	---
SE	8.125	8.541	7.531	---
Method 4 Slope				
γ_{1n}	10.484***	9.019**	3.521	---
SE	2.353	2.583	2.381	---
nWHT x Mthd1Intercept				
γ_{0p}	---	0.365*	0.248	-0.082
SE	---	0.170	0.160	0.155
nWHT x Mthd2 Intercept				
γ_{0q}	-0.365*	---	-0.117	-0.447**
SE	0.170	---	0.166	0.161
nWHT x Mthd3 Intercept				
γ_{0r}	-0.248	0.117	---	-0.331*
SE	0.160	0.166	---	0.150
nWHT x Mthd4 Intercept				
γ_{0s}	0.082	0.447**	0.331*	---
SE	0.155	0.161	0.150	---
FRP x Mthd1 Intercept				
γ_{0t}	---	-0.371	-0.513*	-0.195
SE	---	0.221	0.201	0.195
FRP x Mthd2 Intercept				
γ_{0u}	0.371	---	-0.142	0.176
SE	0.221	---	0.212	0.207
FRP x Mthd3 Intercept				
γ_{0v}	0.513*	0.142	---	0.318
SE	0.201	0.212	---	0.185
FRP x Mthd4 Intercept				
γ_{0w}	0.195	-0.176	-0.318	---
SE	0.195	0.207	0.185	---

*$p < .05$, **$p < .01$, ***$p < .001$.

Chapter Five

Discussion

5.1 The Effects of Score Transformation Methods

Using relative school rankings calculated from percent proficiency as the benchmark for comparison, it does appear that the relative rankings calculated from each of the other three score transformation methods are different. In the multi-level regression of rankings over time on score transformation methods, relative rankings produced by grade-normed proficiency, grade-normed scale scores, and proficiency-centered grade-normed scale scores all have a significantly larger intercept indicating that there are fewer ties in the rankings produced using each of these more complicated score transformations than in the rankings produced through ordering percent proficiency. This significantly larger intercept is, practically speaking, identical for all three score transformations. In the original rankings, ties were handled by assigning consecutive integer values to unique values of test score aggregations. Intuitively then, these results make sense as percent proficiency was a ratio of two whole numbers (the number of students meeting or exceeding standards over the number of students tested) regardless of test, and all other score transformation methods involved norming student- or school-level results using different decimal approximations of means and standard deviations for each grade, subject, and test form. Having more ties in a ranking generated through a simpler calculation seems reasonable. In each set of rankings, the spread of rankings stays the same over time.

Less intuitive to interpret are the differences in the absolute changes in the rankings produced using the different score transformations. The intercept and slope for the changes in rankings produced through percent proficiency indicate that in each consecutive year a school's ranking changes an average of approximately 118 places either up or down. The negative slope indicates that schools' rankings become more stable over time. On average, a school changes nine fewer places, either up or down, each consecutive year. This may be related to the change in math standards for grades 3-8 in 2011. Some stabilization in the scores on the math tests would be expected during the years following a change in standards being assessed. Despite the greater spread of rankings produced using grade-normed percent proficiency, the absolute change in these rankings is not significantly different from that in the percent proficiency rankings. The changes in rankings produced using grade-normed scale scores and proficiency-centered grade-normed scale scores show a statistically significantly ($p < .001$) smaller average change in rankings for a school (an annual average of approximately 109 and 103 places respectively). The average annual change in rank of a school in each of these two sets stays approximately the same over time as indicated by the very small slope coefficients (-2.04 and 1.39, respectively). Essentially, ranks produced through transformations of scale scores show less volatility over time than ranks produced using transformations of proficiency.

In an accountability system such as Minnesota's Multiple Measurements Rating (MMR) that identifies the bottom 5% of schools as priority and the top 15% of schools as reward, the average annual change in rank produced via a transformation of proficiency is

equivalent to moving up or down approximately 8%, whereas the average annual change in rank produced via a transformation of scale scores is equivalent to moving up or down approximately 7%. Though this is a small difference, a relative ranking system with greater movement in ranks may be more conducive to giving more schools a chance to be recognized for their successes.

5.2 The Effects of School Demographics

Considered separately, ranks produced using each of the four score transformation methods show different relationships with school-level demographic predictors. Not surprisingly, ranks produced by all four score transformations have a significantly positive relationship with the percent of a school's testing population identified as qualifying for free or reduced price lunch. For every percent increase in qualifying students the school will, on average, have an increase in rank value of between eleven and fifteen, corresponding to be being between eleven and fifteen places lower on a ranked list. Though this relationship is both statistically and practically significant, it is not alarming. The relationship makes sense in the context of the original purpose of the Elementary and Secondary Education Act (ESEA) and in terms of the achievement gap between students coming from different economic backgrounds that the exams first mandated under the 1994 reauthorization of ESEA, the Improving America's School Act, were supposed to measure.

The first model fitted used the percent of students identified as belonging to any racial/ethnic group other than white as a covariate. These analyses showed that only ranks produced using grade-normed scale scores were dependent on the percent of students at a

school who were not identified as white. However, perhaps counterintuitively given the research on achievement gaps, each percent increase in non-white students resulted in a decrease in rank value of between one and two places. Because of the reverse nature of rankings, a rank of one being higher than a rank of three, this translates to an expected increase in rank of one-to two places for each percent increase in non-white students. That is, schools with more diversity were actually likely to be ranked higher than those with higher percentages of white students. Ranks produced using proficiency and proficiency-centered grade-normed scale scores similarly predicted a decrease in rank value of between one and two for every increase in percent of students identified as having limited English proficiency.

As mentioned earlier, there were also statistically significant relationships between school-level demographics and the average absolute change in rankings between consecutive years. Rankings produced using all four transformation methods show that ranks are more stable over time (change fewer places up or down) for schools with higher percentages of students who are identified as non-white. Using rankings produced from grade-normed scale scores, schools with higher percentages of students identified as qualifying for free or reduced price lunch see, on average, more change in ranks between consecutive years. Lastly, rankings produced from grade-normed proficiency also show significantly more change in ranks for schools with higher percentages of students identified as having limited English proficiency. The most concerning and least defensible school demographic appearing to have a significant relationship with a school's change in rank is the percent of students identified as belonging to any of the

non-white racial/ethnic categories. A student's racial/ethnic identify is a permanent characteristic while a student's English language skills and/or economic situation can change over time. Having a relative school ranking system that is dependent on the racial/ethnic demographics of the students a school serves may motivate schools to try to change who is served rather than trying to change what programs and supports can be added or modified to support the students already being served.

To delve more deeply into the relationship between school ranks and student racial/ethnic backgrounds, four additional models using the percent of students at each school identified as each of the four non-white racial/ethnic categories as predictors were fitted. There was a significant relationship between rank and the percent of students identified as Hispanic for ranks produced using proficiency and grade-normed proficiency. On average, for each percentage increase in Hispanic students served by a school a between two and three point increase in rank value was observed. That is, schools serving larger populations of Hispanic students were likely to be ranked lower (have a higher rank value). For ranks produced using grade-normed scale scores, schools serving larger populations of students identified as Asian/Pacific Islander or black were likely to have a smaller rank value and be ranked higher relative to other schools. Only ranks produced using proficiency-centered grade-normed scale scores showed no significant relationships to any of the school-level racial/ethnic covariates.

Absolute changes in rankings produced using all four score transformation methods were significantly lower (less movement either up or down) for schools serving higher percentages of students identified as Asian/Pacific Islander, Hispanic, or black.

Additionally, for ranks produced using grade-normed proficiency and grade-normed scale scores, absolute changes in rankings were significantly lower for schools serving higher percentages of students identified as American Indian/Alaskan Native.

5.3 The Effects of Interactions of School Demographics and Score Transformation Methods

When school-level demographic variables, the method of score transformation, and the interactions of these characteristics were all included in the model for predicting rank over time, interesting results emerged. The first of these results was that through systematic trimming of the full model, eliminating covariates that were not significant in predicting ranks calculated using any of the score transformation methods, only two school-level demographic predictors remained in the model for rank: the percent of students identified as non-white and the percent of students identified as receiving free or reduced price lunch. In the trimmed models predicting absolute change in rank over time, all four school-level demographic predictors stayed in the model: the percent of students identified as non-white, the percent of students identified as receiving for free or reduced price lunch, the percent of students identified as having limited English proficiency, and the school-level percent mobility.

In the models using rank over time as the outcome variable, model coefficients indicated that ranks produced using proficiency and proficiency centered-grade normed scale scores were not significantly different from each other. Also, ranks produced using grade-normed proficiency and grade-normed scale scores were not significantly different from each other. Ranks produced using transformations of scale scores were significantly

related to the percent of the students at a school identified as non-white, while ranks produced using transformations of proficiency were not. School ranks produced using all score transformation methods other than proficiency, were significantly more related to the percent of students receiving free or reduced-price lunch than the school ranks produced through proficiency were. Lastly, school ranks produced using grade-normed scale scores were significantly more related to the percentage of students identified as non-white than those produced using proficiency or grade-normed proficiency. All of these model coefficients support the conclusion that school ranks produced through proficiency – the simplest score transformation method – are the least dependent on school demographics.

In the models using absolute change in ranks over time as the outcome variable, two model coefficients were notable. Changes in school ranks produced using grade-normed proficiency were significantly more related to the percent of students identified as non-white than those found using proficiency and absolute changes in school ranks produced using grade-normed scale scores were significantly more related to the percent of students identified as receiving free or reduced price lunch than those found using proficiency.

5.4 Choosing the Optimal Ranking System

Earlier, it was stated that an optimal ranking system would be using the simplest score transformation related to the fewest number of school demographics. Results of all three sets of analyses point to the ranking system related to the fewest number of school demographics as the one that uses the simplest score transformation. Relative school

rankings based on percent proficiency (the number of students who are proficient divided by the number of students taking the test) are only significantly dependent on the percent of students at each school who are receiving free or reduced price lunch. How much those school ranks change over time is significantly related to the percent of students identified as either non-white or having limited English proficiency and how those changes increase or stabilize over time is related to the percent of students receiving free or reduced price lunch and the school-level mobility rate.

One implication of this finding is that the more complicated score transformation methods employed by process such as creating the quality index for charter schools (grade-normed percent proficiency), or for selecting schools to receive the National Blue Ribbon Schools award (grade-normed scale scores), may not only be unnecessarily complex but may also be inadvertently ranking schools so that some school receive awards and opportunities based on who they serve rather than on school quality. Another implication of this finding is that organizations such as the Minnesota Department of Education may be guilty of creating avoidable barriers to wide-spread public understanding of their ranking systems.

5.5 Caveats, Considerations, and Future Research

Of course, one must remember that the entire premise of this study is based on the assumption that the tests used to measure mastery of state standards are also a measure of school quality. This is not an assumption that is unique to this paper. Federal law requires the use of performance on these tests to be used as a measure of school quality (USDOE, 2010). However, federal law also allows for the inclusion of other measures, such as

growth and graduation rate, to be a part of the measure of a school's quality or success with serving students. Any of the other measures used to quantify a school's relative quality should also be analyzed to determine its level of dependence on school or student demographics.

Another big assumption made in this study is that Minnesota state accountability exams (MCAs, MODs, and MTASs) measure students from all backgrounds the same way. It is unclear if analysis has been performed on potential differential item functioning between groups of students. If such differential item functioning does exist, then any aggregation of scores, no matter the complexity of score transformation, will inherently be dependent on student demographics. As it stands, it is possible that the ESEA requirement that all students be held to the same standards means that even if differential item functioning were known to exist, no adjustments could be made to test scores to account for it (IASA of 1994, 108 Stat. 3524).

A limitation of this study is that only score transformation methods already in use by the Minnesota Department of Education were compared. Methods of score transformation and/or aggregation used in other states or in fields outside of education should also be examined for dependence on school demographic characteristics. For example, it is possible that forecasting techniques used in sports or economics could be used to project what a school's future performance should be and then to assess the extent to which the school meets that expectation. Additionally, it would be interesting to consider how value-added models - models that consider a student's previous performance on a test as the baseline for current and future performance – might be

applied to Minnesota's standardized tests. Without being able to compare scores directly across years of testing, some other type of score transformation may need to be used to make the changes in an individual student's score across years meaningful (MDE, 2012b). There may also be applications of value-added modeling at the classroom or school level that could improve a system of relative ranking. This line of inquiry may become especially relevant as more schools and districts tie standardized test results directly to teacher and program evaluations (Hanushek & Rivkin, 2010; Harris, 2011).

As the ESEA is reauthorized and amended, the needs for ranking schools and the acceptable methods for doing so will likely evolve as well. Currently, no monetary incentives or punishments are attached to school rankings required by the federal law. However, both monetary and recognition-based incentives are awarded through related ranking systems and the general national desire to use individual test scores as aggregate measures of teacher skill and school quality is growing. For now, the results of this study indicate that a tendency toward parsimony through the use of the simplest test score transformation will result in an optimal aggregate measure for ordering schools.

References

Abedi, J. (2002). Standardized achievement tests and English language Learners: Psychometric issues. *Educational Assessment, 8,* 231-257.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33,* 4-14.

Anonymous. (2010, June 21). What makes a good school? *New Statesman, 139* (2006), 18.

Beard, T. R., & Caudill, S. B. (2009). Who's number one? – ranking college football teams for the 2003 season. *Applied Economics, 41,* 307-310. doi:10.1080/00036840601007245

Beckman, T. O., Messersmith, K., Shepard, J., & Cates, B. (2012). Ethnicity, language and poverty predicting scores on the Nebraska state accountability reading test. *International Journal of Psychology: A Biopsychosocial Approach, 11,* 31-47. doi:10.7220/1941-7233.11.2

Borgatta, E. F. & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods and Research, 9,* 147-160. doi:10.1177/004912418000900202

Demie, F. (2002). Pupil mobility and educational achievement in schools: An empirical analysis. *Educational Research, 44,* 197-215. doi:10.1080/00131880210135304

Elementary and Secondary Education Act of 1965, Pub. L. No. 89-10, § 79 Stat. 27-58.

Fischer, S. & Stoddard, C. (2013). The academic achievement of American Indians. *Economics of Education Review, 36,* 135-152. doi:10.1016/j.econedurev.2013.05.005

Garcia, D. R. (2008). Mixed messages: American Indian achievement before and since the implementation of no child left behind. *Journal of American Indian Education, 47,* 136-154.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research, 45*, 43-57. doi: 10.3102/00346543045001043

Goals 2000: Educate America Act, Pub. L. No. 103-227, § 108 Stat. 125-280.

Gottfried, A. (2014). The achievement effects of tardy classmates: Evidence in urban elementary schools. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 25,* 3-28. doi:10.1080/09243453.2012.728135

Hanushek, E. A. and Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review, 100*, 267-271. doi:10.1257/aer.100.2.267

Harris, D. N. (2011). Value-added measures and the future of educational accountability. *Science, 333*, 826-827. doi:10.1126/science.1193793

Improving America's Schools Act of 1994, Pub. L. No. 103-382, § 108 Stat. 3518-4062.

Kendal, M. G. (1945). The treatment of ties in ranking problems. *Biometrika, 33,* 239-251. doi:10.2307/2332303

Kolen, M. J. and Lee, W. C. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice, 30*(2), 15-24. doi:10.1111/j.1745-3992.2011.00201.x

Ladd, H. F. (2012). Presidential address: Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management, 31*, 203-227. doi:10.1002/pam.21615

Lonetree, A. (2013, September 7). Local students fare well in 2013 state standardized tests. *Star Tribune.* Retrieved from http://www.startribune.com/local/east/222820471.html

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19-40. doi:10.1037//1082-989X.7.1.19

Madrid, E. M. (2011). The Latino achievement gap. *Multicultural Education, 19*(3), 7-12.

Magan, C., Webster, M.J., & Koumpilova, M. (2013, August 26). Minnesota math, reading scores slip, but science proficiency up slightly. *Pioneer Press.* Retrieved from http://www.twincities.com/localnews/ci_23950588/mca-proficiency-scores-fall-math-reading?IADID=Search-www.twincities.com-www.twincities.com

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012) TIMSS 2011 international results in science. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-343.

McKay, R. E. (1965). The Elementary and Secondary Education Act of 1965: the President's program: 'a new commitment to quality and equality in education'. *The Phi Delta Kappan, 46,* 427-429.

Mehana, M., Reynolds, A. J. (2004). School mobility and achievement: A meta-analysis. *Children and Youth Services Review, 26,* 93-119. doi:10.1016/j.childyouth.2003.11.004

Minnesota Department of Education. (n.d.-a). [Graphic and numerical test results]. *Minnesota Report Card.* Retrieved from http://rc.education.state.mn.us/#

Minnesota Department of Education. (n.d.-b). *Interim high-quality charter school method and process summary*. Retrieved from http://education.state.mn.us/MDE/StuSuc/EnrollChoice/CharterSch/index.html

Minnesota Department of Education. (2011, September 20). *Attachment 15: Functional requirements for the 2011 NCLB AYP calculations.* Retrieved from http://education.state.mn.us/MDE/SchSup/ESEA/FedAcc/ESEAFlexReq/index.html

Minnesota Department of Education. (2012a, February 7). *ESEA flexibility request.* Retrieved from https://www2.ed.gov/policy/eseaflex/approved-requests/mn.pdf

Minnesota Department of Education. (2012b). *Interpretive guide for Minnesota assessment reports 2012-2013.* Retrieved from http://education.state.mn.us/MDE/EdExc/Testing/index.html

Minnesota Department of Education (2014). *Functional Requirements for the Minnesota Multiple Measurement System 2014.* Retrieved from http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

Morales, M. C. & Saenz, R. (2007). Correlates of Mexican American students' standardized tests scores: An integrated model approach. *Hispanic Journal of Behavioral Sciences, 29,* 349-365. doi:10.1177/0739986307302176

Morse, R. (2013, April 22). *Frequently asked questions: Best high schools rankings.* Retrieved from http://www.usnews.com/education/high-schools/articles/2013/04/22/frequently-asked-questions-best-high-schools-rankings-2

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012) TIMSS 2011 international results in mathematics. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Center for Education Statistics. (2010, December). *Statistical methods for protecting personally identifiable information in aggregate reporting* (Brief No. 3). Retrieved from http://nces.ed.gov/pubs2011/2011603.pdf

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425-2094.

Nobles, J., Alter J., Cherin, D. Connelly, C., Hauer, J., Jacobsen, D., & Yunker, J. (2008, June). Evaluation report: Charter schools. Retrieved from http://www.auditor.leg.state.mn.us/ped/2008/charterschools.htm

OECD (2013a). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

OECD (2013b). *PISA 2012 Results in focus.* Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

R Development Core Team (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raudenbush, S. W., Bryk, A. S,, & Congdon, R. (2004). HLM 6 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Rowley, R. L. & Wright, D. W. (2011). No "white" child left behind: The academic achievement gap between black and white students. *The Journal of Negro Education, 80*(2), 93-107. doi:10.2307/41341113

Seife, C. (2010). *Proofiness: How you're being fooled by the numbers.* New York: Penguin Group.

Strand, S. (2014). School effects and ethnic, gender and socio-economic gaps in educational achievement at age 11. *Oxford Review of Education, 40*, 223-245. doi:10.1080/03054985.2014.891980

State's Impact on Federal Education Policy Project. (2009). *Federal education policy and the states, 1945-2009: A brief synopsis.* Retrieved from http://nysa32.nysed.gov/edpolicy/altformats/ed_background_overview_essay.pdf

Stevens, S. S. (Ed.). (1966). *Handbook of experimental psychology.* New York, NY: Wiley.

Student. (1921). An experimental determination of the probable error of Dr Spearman's correlation coefficients. *Biometrika, 13,* 263-282. doi:10.2307/2331754

Thorndicke, R. M., and Thordike-Christ, T. (2010). *Measurement and evaluation in psychology and education.* (8th ed.). Boston, MA: Pearson Education, Inc.

Turkan, S. & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education, 34,* 2343-2369. doi:10.1080/09500693.2012.705046

U.S. Department of Education. (n.d.). *National blue ribbon schools program.* Retrieved from http://www2.ed.gov/programs/nclbbrs/index.html

U.S. Department of Education (2012, June 7). *ESEA flexibility*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2014). *The national assessment of educational progress (NAEP).* Retrieved from http://nces.ed.gov/nationsreportcard/

U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2010). *ESEA Blueprint for Reform,* Washington, D.C.

van Dulmen, M. H. M., & Egeland, B. (2011). Analyzing multiple informant data on child and adolescent behavior problems: Predictive validity and comparison of aggregation procedures. *International Journal of Behavioral Development, 35,* 84-92. doi:10.1177/0165025410392112

Wright, D. (1999). Student Mobility: A negligible and confounded influence on student achievement. *The Journal of Educational Research, 92*, 347-353.

Young, J. W., Cho, Y., Guangming, L., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*, 170-192.

Zhu, W. (2013). Test equating: What, why how? *Research Quarterly for Exercise and Sport, 69*, 11-23. doi:10.1080/02701367.1998.10607662