Validation of Ohio's Proposed Reforms for K-12 Accountability Systems

Dissertation

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

By

Kara Lynn Breda

Graduate Program in Educational Studies

The Ohio State University

2018

Dissertation Committee

Jerome V. D'Agostino, Ph.D., Advisor

Dorinda J. Gallant, Ph.D.

P. Cristian Gugiu, Ph.D.

Abstract

Ohio House Bill 2 recently passed and included a provision for the study of the Similar Schools Measure (SSM) which is in use by the California Department of Education. The SSM is a weighted multiple regression model that contains variables to determine if schools are outperforming what is expected based on their demographic makeup. This provision was in response to the legislative push for a model that includes demographic variables and a proposal to enact a model created by the Ohio Coalition for Quality Education (OCQE). The OCQE model is a multiple regression model that is very similar to the SSM but is modest in that it uses fewer variables to create school achievement predictions and is not weighted.

The purpose of the proposed study is to validate and compare the accountability models being considered by the Ohio legislature. The models were replicated using the Ohio Department of Education's (ODE) data of school-level demographics and achievement scores as calculated on ODE's school report cards. The analyses included from 3,012 schools for the 2011-2012, 2012-2013, and 2013-2014 school years. Model fit and statistical assumptions are examined to ascertain the viability of the models of interest. Comparisons of the models and were conducted using descriptive statistics and correlation analyses. Additionally, subgroup difference for charter schools, socioeconomic levels, and typology of schools were conducted. In order to evaluate the

reliability of the models, a generalizability study was used to determine sources of systematic variability.

Results indicated that model outcomes and significance were similar to each other and to the current achievement measure. The models and measures were highly correlated but lesser strength correlations were found within specific subgroupings. Across years, consistent predicted scoring and subgroup differences were found. Although the models showed statistical significance and consistency, the suitability of the models is in question due to assumption violations of heteroscedasticity, linearity, and multicollinearity. Reliability analyses of a generalizability study and Cronbach's alpha illustrated the consistency of the models across years but was unable to designate a more reliable model due to model outcome similarities.

Overall, the proposed models aligned with previous research as to the effects of demographic variables on achievement. However, when examining subgroup differences, both the OCQE model and SSM did not provide outcomes that heavily fluctuated from the current achievement measure. The reliability of the models was found to be satisfactory but the appropriateness of the use of multiple regression without transformations is in question. Further research is needed to correct assumption violations.

Acknowledgments

I would like to sincerely thank my committee members, Dr. D'Agostine, Dr. Gugiu, and Dr. Gallant, for their advice and assistance in this process.

Thank you to my husband for being incredible supportive throughout the years.

Vita

2008……B. A. Psychology, Michigan State University

2010……M. A. Industrial-Organizational Psychology, Xavier University

Fields of Study

Major Field: Education Studies

Major Field: Industrial-Organizational Psychology

Table of Contents

List of Tables

Chapter 1. Introduction

Study Background

The United States education system is undergoing changes in the way we look at student achievement and school performance. Based on reforms beginning in the 1980s, states have introduced accountability systems that use quantitative and qualitative measures to show comparisons among institutions (Wellmam, 2001). The bulk of the measures that make up accountability systems are centered on student assessments. The reliance on student assessments comes from the premise that, "student achievement will improve when individual schools are held accountable through the mechanisms of establishing standards and evaluating performance of students, and subgroups of students, on annual standardized tests." (Itkonen & Jahnukainen, 2007).

Each state is able to develop their own policies for accountability systems which allows them to decide the standards, tests, and ratings that schools are measured on (Lee, 2010). This flexibility has made for a multitude of ratings and rankings in which schools are judged and, in the case of charter/community schools, potentially closed. It is important to recognize that the accountability systems adopted by states are classified as consequential accountability systems. Consequential accountability systems have state implemented sanctions or rewards for schools based off of school and district performance on accountability systems. These high-stakes accountability systems spread

in the 1990s with the introduction of No Child Left Behind (NCLB), (Kress, Zechmann, & Schmitten, 2011).

NCLB is a federally mandated act that requires schools to demonstrate "Adequate Yearly Progress" (AYP) in order to receive federal funds (Hanley, Roehrig, & Canto, 2015). Student scores on state assessments in math and reading, participations in the assessments, and high schools graduation rates or Kindergarten through 8th grade attendance serve as measures for meeting AYP.

Current accountability systems include measures such as status models, growth models, and gap closing. Linn (2008) provides a summary of the various types of measures that are used in test-based accountability systems. Status models are based on "achievement targets" of annual state assessments. Growth models focus on the gains or losses in achievement instead of single year achievement scores. Longitudinal tracking of individual achievement is required for growth models to measure learning over time. Gap closing refers to the goal of minimizing achievement differences between identified subgroups and overall student populations.

Within each of the above measures there can be methodological variations and each state combines or prioritizes the measures differently. While states have flexibility on the structure of their accountability systems, they still need to incorporate federal measures such as Adequate Yearly Progress. This has led to confusion on a school's performance as it may appear to perform at a high level on state measures but not meet federal measures (Linn, 2008). It can also be difficult for parents and educators to ascertain how well a school performs since a school can display poor achievement scores

on a status measure but show adequate progress based on growth measures or vice versa. Growth models are seen as more robust as they take into account multiple years of data, typically use complex statistical techniques, and usually include a covariate of previous achievement scores (Castellano & Ho, 2013). Status models show point-in-time learning achievement whereas growth models show a culmination of student learning. However, both status and growth models can vary drastically depending on what variables, if any, are used for controls (such as mobility or disability status) and what procedures are used to get to the final metric. In the case of Ohio's growth measure, a proprietary technique and software are used that has not been released to the public so only the general layout of the formula is known but the status measure's formula is publicly available.

Differences occur not only on the aggregating measures but also on how much students are determined to have proven content knowledge. Even though gains in reading and math have been shown through state mandated assessments, results on the National Assessment of Educational Progress (NAEP) do not reveal the same gains (Lee, 2010).

Due to the significance of accountability systems it is important to fairly and validly evaluate schools. A student's performance is not solely determined by school or teacher influences. Educationally irrelevant variables such as socioeconomic status, disabilities status, or student mobility can impact student achievement and therefore impact accountability measures (Lee, 2007). Identifying underlying factors of educational performance, and including those in accountability systems, may help to understand and better rate the performance of schools.

Study Purpose

Ohio House Bill 2 recently passed and included a provision for the study of the Similar Students Measure (SSM) model which is in use by the California Department of Education. The SSM is a weighted multiple regression model that contains variables to determine if schools are outperforming what is expected based on their demographic makeup. This provision was in response to the legislative push for a model that includes demographic variables and a proposal to enact a model created by the Ohio Coalition for Quality Education (OCQE). The OCQE model is a multiple regression model that is very similar to the SSM but is modest in that it uses fewer variables to create school achievement predictions and is unweighted. Both the OCQE model and SSM are status models and thus focus on point-in-time achievement for schools. The purpose of the proposed study is to validate and compare the status accountability models being considered by the Ohio legislature that could potentially be used to evaluate schools.

Research Questions

Although the SSM is being studied by the Ohio Department of Education, the OCQE model would most likely be implemented if the legislation is successful. It is important to determine if there are significant differences between the SSM, OCQE model, and the current accepted achievement measure. It is also important to determine if the models proposed are acceptable models for the data that will be used. Model fit and statistical assumptions will be examined to ascertain the viability of the models of interest. In order to assess the reliability of the models, a generalizability study will be

used to determine sources of systematic variability. The proposed study will seek to answer the following primary questions:

1. To what degree do school-level demographic characteristics influence school achievement scores in Ohio? Specifically, how much do demographic characteristics account for variability in the OCQE model and SSM?
2. What are the relationships between the OCQE model, SSM, and the Performance Index rating?
3. Are the OCQE model and SSM adequate representations of the given data?
4. Utilizing a generalizability study for the OCQE model, SSM model, and the Performance Index, which model is more reliable for making decisions based on school performance?

Organization of the Dissertation

Chapter 1 of the dissertation provides an introduction, study background, and study purpose. Chapter 2 gives a detailed literature reviews. The literature review includes research on accountability systems, examples of state accountability measures, common factors influencing student achievement, and an explanation of generalizability theory. Chapter 3 provides the methodology for the present study. Sections include data sources used, explanations of Performance Index, the Ohio Coalition for Quality Education model, Similar Students Measure, and the procedure for the study. Chapter 4

details the results on all analyses for the study. Finally, Chapter 5 provides the discussion, conclusion, limitation, and implications for future research.

Chapter 2. Literature Review

Accountability Systems in the United States

Assessing Accountability

Educational accountability systems typically comprise of achievement standards for student performance measures that are attached to a system of consequences based on the ability to meet the given standards (Gándara & Randall, 2015). The measures that are rated can include attendance rate, graduation rate, and student performance on state mandated testing (Kelly & Orris, 2011). State mandated testing has taken the forefront in accountability after the No Child Left Behind Act (NCLB) required all schools to give reading and math assessments to 3rd through 8th graders annually (Derthick & Dunn, 2009). Proficiency goals on state assessments were created as academic standards that included goals for subgroups such as English language learners and students with disabilities.

For the United States, Conley (2015) describes the current methodology of educational accountability systems. Simply put, only specific subject areas are indicators of a student's achievement for the "knowledge, skills, and capabilities" of those attending public schools. With budgets in mind, many policymakers look to minimize testing costs but rely on them to draw wide conclusions and make decisions on the effectiveness of schools and schools systems. Focus is now pivoting towards college and career readiness which has led educators and parents to think more critically about how student learning and school effectiveness is assessed.

Research has looked into the relationship between high-stakes testing accountability systems and student achievement. Marchant, Paulson, and Shunk (2006)

examined how NAEP testing results were related to high-stakes testing states. The authors found that comparisons of NAEP scores over multiple years showed being in a high-stakes testing state led to slightly lower achievement scores for reading and science but showed slightly higher growth scores. However, once demographic factors were controlled for, significant relationships based on high-stake testing were no longer there.

Taking an international approach, effects of accountability systems on science achievement were investigated by Gándara & Randall (2015). Four countries, including the United States, Australia, Korea, and Portugal were included in the analysis. The countries were analyzed for differences in practices of accountability systems and their relationship to science achievement based on PISA results. The Programme for International Student Assessment (PISA) is an "international assessment program that measures the skills and knowledge of 15-year-olds across three domains: literacy, mathematics, and science" that takes occurs every three years in numerous countries. The United States was found to have higher-level and more influential accountability practices than the other countries. The relationship between accountability practices and science achievement was found to be mainly negative and small but the effects of student-level socioeconomic status on achievement far outweighed any other effects.

A study by Jennings and Sohn (2014) analyzed the effects of accountability systems on high and low performing students by examining data from the year before and year after the implementation of NCLB in Houston, Texas. Interestingly, the findings indicated that inequalities between high and low performing students increased for math but decreased for reading. The results were admittedly puzzling by the authors and only

gave a short-term view on potential consequences. This study highlights how difficult it can be to come to conclusions for complex systems.

Accountability policies focused on evaluations of teachers have also been examined for effects on student achievement. Research has found that having state evaluation models for teachers based on student achievement increase reading proficiency rates but no relationship was found for math proficiency for 8th grade students (Alexander, Jang, & Kankane, 2017).

Based on current research it is difficult to ascertain what, if any, effects high-stakes accountability environments have on student achievement. The research presented shows how any found relationships appear to be small, inconsistent, and can be shadowed by the effects of demographic factors. Future research will need to determine if the risks and rewards of accountability systems are contributing to desired outcomes.

Accountability system policies and effects have been researched but few studies have been conducted that evaluate the models or measures accountability systems are based on. Most of the research on measures within accountability systems focuses on growth models. Growth models can be mainly categorized as growth description, growth prediction, and value-added (Castellano & Ho, 2013). Growth description creates a magnitude of growth for an individual or group, growth prediction estimates future scores given a student's current and past achievements, and value-added provides links between gains and educators or schools. Anderman, Anderman, Yough, and Gimbert (2010) summarized debates on value-added growth models. The authors found that researchers disagreed on methodological issues such as whether to control for demographic

characteristics in the model, if missing data should be treated as random, and if the assessments that are given to students can adequately measure growth over consecutive grades. When comparing different types of growth models on school ratings, Goldschmidt, Choi, and Beaudoin (2012) found that models performed similarly if they were of the same category but had wide variations if demographic characteristics were used. The authors outlined that various growth models answer various questions so education departments have to decided what type of inferences they want to make and choose a model accordingly. Studies on status models are limited even though many states use some form of yearly achievement measure. The present study will contribute to filling the research gap on status models.

State Accountability. Each state in America has their own unique accountability systems for K-12 education. While there are federal guidelines that need to be met, every state has latitude to measure and hold schools accountable as they see fit. The state education systems tend to have similar measures or indexes, such as growth and performance, for schools but do not have the same calculations to get to the end results. For illustrative purposes, Ohio, Michigan, and Florida's K-12 accountability systems will be examined.

Ohio's accountability measures. Ohio's current system includes district and school building report cards made up of 6 components (Ohio Department of Education, 2015a). These six components include an Achievement rating, Progress rating, Gap Closing rating, Graduation Rate, K-3 Literacy rating, and a Prepared for Success rating. Each component is calculated into a percentage and then converted to points. If a

component has sub-measures, then each sub-measure is weighted to create an overall point value. The points are then equated to an A, B, C, D, or F grade for each component.

The Achievement component is made up of the sub-measures Performance Index and Indicators Met. The Performance Index measures the achievement of every student, not just whether he or she reaches "proficient." Schools receive points for every student's level of achievement on the state tests. The higher the student's level, the more points the school earns towards its index. The levels range from 1 to 5, with 3 being considered proficient. The Indicators Met rating is measured by the percentage of students who have reached proficient on each state test.

The Progress component includes four sub-measures of Overall, Gifted Students, Students in the Lowest 20% in Achievement, and Students with Disabilities. The Overall sub-measure includes all students and all test subjects. Each sub-measure is the Value-Added score for each grouping which shows the students' growth during the school year based on what a year's worth of growth is estimated to be within Ohio's student population.

The Gap Closing component is comprised of Annual Measurable Objectives (AMOs), which are set performance expectations of sub-groups on English Language Arts assessments, Math assessments, and Graduation rates. Each year the expectations are adjusted and vary between the objectives. For example, the 2016 AMO for English Language Arts was 74.2%, Math was 68.5%, and Graduation rate was 82.8%. The sub-groups considered are All Students, American Indian/Alaskan Native, Asian/Pacific Islander, Black/non-Hispanic, Hispanic, Multiracial, White/non-Hispanic, Economically

Disadvantaged, Students with Disabilities, and Limited English Proficiency students. The percentage of sub-groups that meet the expectations creates the AMO score.

Graduation Rate is calculated by dividing the number of students who graduate high school in four years or less & five years by the number of students who form the adjusted cohort for the graduating class. The adjusted cohort is any student who started 9th grade for the first time in a given year. For example, any student who started 9th grade in the 2010-2011 academic year would be expected to graduate in 2013-2014 academic year, or sooner, to be counted in the four-year graduation rate. The same student would need to graduate in 2014-2015 to be counted in the five-year graduation rate.

The K-3 Literacy component uses results from reading diagnostic assessments given to all students in kindergarten through the beginning of the year of 3rd grade. Points are given by having students move from "not on-track" to "on-track" from one year to the next.

The final component is Prepared for Success. The component includes the students in the adjusted cohort from the graduation rate and gives one point to each student who earns a remediation-free score on all parts of the ACT or SAT, earns an honors diploma, or earns an industry recognized credential (such as an AutoCad User credential or American Welding Society – Certified Welder credential). Bonus points are given for students who also earn a 3 or higher on AP exams, 4 or higher on IB exam, or earn at least 3 college credits before graduating.

Due to each school's population being different, it is possible that some schools will not have grades for some of the components. A high school would not have a K-3

Literacy grade because they do not have kindergarteners through third graders and a school without a Hispanic population would not have an AMO for that subgroup. If components, sub-measures, or sub-groups are not available for the school, the grading calculations simply adjust so that there is more weight given to the existing categories. The above version of Ohio's school report card began in 2013 and has slowly rolled out official letter grades for each of the components. By 2018 the state plans to implement letter grades for all components and an overall school letter grade.

Michigan's accountability measures. Michigan's school accountability score cards include 5 components made up of participation rates on state assessments, proficiency rates on state assessments, attendance or graduation or rates, educator evaluations, and compliance factors (Michigan Department of Education, 2014). Participation rates are calculated by dividing the number of students with valid assessments by the number of students enrolled at the school.

The proficiency rate component has two sub-components consisting of proficiency targets and proficiency growth. Proficiency targets are uniquely set for each school and district based on the previous year's percent proficient. Proficiency growth is used to supplement any non-proficient scores by allowing those students who show improvement from the previous year to be counted as a proficient score.

Schools that do not receive graduation rates are measured on attendance rates. Attendance rates are determined by taking the total days students attended a school and dividing that quantity by the total days of possible attendance for the school. All schools must meet an attendance rate target of 90%. For the schools that have graduation rates,

calculations are based on four-, five- and six-year graduation rates. If all of the graduation rates are below 80% then the school must meet improvement targets to satisfy the requirement.

Educator evaluations have two sub-components, which are data reporting requirements for Effectiveness Labels and Teacher Student Data Link (TSDL). Teachers are rated as either Highly Effective, Minimally Effective or Ineffective, with potential dismissal for being rated Ineffective for three consecutive years. This Effectiveness Label has to be submitted for 100% of teachers to meet the report card target. The TSDL is data submitted that links students to the teachers who provided them with instruction throughout the year. At least 95% of teacher and students must be linked in order to meet the report card target.

The final component, compliance factors, requires schools to complete a school improvement plan and a school accreditation report. Each report is specific to the school and both need to be completed to fulfill the requirement.

Michigan also classifies schools as Reward schools if they fall into the highest performing schools. As a subset of Reward schools, if a school outperforms expectations based on demographic variables then they are considered “Beating the Odds” (Michigan Department of Education, 2013). A Beating the Odds school is one that either outperforms their predicted performance or outperforms demographically similar schools. A school’s predicted performance is based on a multiple regression equation where the school rank is the outcome variable. The predictor variables include percent economically disadvantaged, percent students with disabilities, percent English language learners, and

percent minority. A 95% confidence interval is constructed around each predicted rank using the standard error of prediction and if the school is above the upper bound limit, then they are considered Beating the Odds. Outperforming demographically similar schools is determined by grouping schools into sets of 30 using a weighted standardized Euclidean distance. The equation includes the following school data: grade levels served, total enrollment, state foundation allowance, percent economically disadvantaged, percent with disabilities, percent English language learners, percent minority, and also tracks whether the school is over 80% students with disabilities. If a school has the highest rank within their set and had a statistically significant higher ranking than the group, at the α=.001 level, then the school is considered Beating the Odds.

Florida's accountability measures. Florida's accountability system includes up to 11 components and sub-components for each school. The components are grouped by achievement, learning gains, middle school acceleration, college and career acceleration, and graduation rate. Each component is worth up to 100 points which are added together, divided by the total possible points for that school and then made into A-F letter grades based on the percentage of points earned (Florida Department of Education, 2014). Schools must also test at least 95% of their students, have sufficient data for at least one component, and have more than 10 eligible students in a given component in order to receive a letter grade.

The achievement component in Florida's score card includes four sub-components of English language arts, mathematics, science, and social studies. Each sub-component measures the percentage of full-year enrolled students who scored at an

achievement level of 3 or higher on standardized assessments, which signifies a passing score. The assessments span grades 3 through 12 and have scoring levels of 1 through 5.

The learning gains component includes four sub-components of learning gains in English language arts, mathematics, the lowest performing 25% in English language arts, and the lowest performing 25% in mathematics. Learning gains are conceptualized as students advancing in scoring levels from the prior year to the current year. In levels 1 and 2 there are sub-levels of Low, Middle, and High that students can move to in order to count as having a learning gain.

Middle school acceleration is the percentage of eligible students who have gained an industry certification or passed an end-of-course state assessment at the high school level. College and career acceleration is the percentage of high school graduates who gained an industry certification, earned a passing score on an accelerated examination, such as an AP or IB exam, or who earned college credit through a dual enrollment course.

Graduation rate is calculated based on a 9th grade adjusted cohort and the percentage of students to graduate within four years. The adjusted cohort is considered the students who began 9th grade for the first time, four academic years prior. Transfer-out students are removed from a school's equation and transfer-in students are added to the year's cohort in which they belonged in their previous school. However, students who transfer to an adult education program or a Department of Juvenile Justice facility will stay a part of their regular school's cohort and calculation.

Florida also uses Value-Added growth modeling as part of their teacher evaluation system, but not as a part of the school grade card. Florida uses a covariate

adjustment model that considers student, classroom, and school covariates. These covariates include two years of prior assessment scores, the number "subject relevant" courses that the student is enrolled in, disability status, English Language Learner status, gifted status, attendance, number of school transitions, an indicator of grade retention, class size, and homogeneity of students' test scores in the given class. The covariates also include the above factors at the school-level (Florida Department of Education, 2014).

The states above, as with each state in the country, determine the types of variables to include in each of their accountability systems. While it is typical for states to look at performance and ratings between demographic groups, not all states directly include those types of variables in their models or calculations. Currently, Ohio does not include demographic data in their Value-Added calculation while Florida does include characteristics such as disability status, number of school transitions, and gifted status. Michigan does not include demographic characteristics in their component grades but does have a regression-based model which uses school-level data on the percent economically disadvantaged, percent minority, percent with disabilities, among others, to determine which schools are outperforming based on the demographic makeup of the school.

Common Factors Influencing Student Achievement

The common methodology of many accountability systems is that schools and teachers and school policies are the drivers of student academic performance. In other words, accountability systems seek to measure schools based on what is the school's control. School and teacher influences on student achievement can be significant

however, much research has shown that student factors are the main drivers. While the present study will focus on the effects of student factors that influence achievement, the next sections will give an overview of recent research into school-based and teacher-based effects.

School-based factors. School environmental, or climate, conditions that have been theorized to influence student outcomes include school size (Schwartz, Stiefel, & Wiswall, 2016), differentiated instruction as a schoolwide approach (Goddard, Goddard, & Kim, 2015), enrollment patterns in "School Choice" areas (Ahn & McEachin, 2017), and building condition (Maxwell, 2016). Wang and Degol (2016) summarized school climate as, "virtually every aspect of the school experience, including the quality of teaching and learning, school community relationships, school organization, and the institutional and structural features of the school environment." The authors reviewed the large quantity of research that examined relationships between school climate and academic outcomes. They observed that the most consistent findings demonstrated high academic achievement in schools where high academic standards were set, stressed commitment to students, exhibited effective leadership, and emphasized mastery goal orientations. However, institutional factors such as school size, type, location, structural features, and socio-economic status were inconsistent in their investigations.

Another lens of school climate is the students' perceptions of the climate. Gietz and McIntosh (2014) studied students' perceptions of safety, inclusion, experience with being bullied/victimization, and clear expectations of behavior. The results indicated that, controlling for school-level poverty, a positive view of the school climate was associated

with academic achievement. Grade level and subject achievement differences were also seen, for instance differences in the highest predictor of achievement. Fourth graders displayed victimization and 7th graders displayed clear expectations of behavior as the highest predictor.

Teacher-based factors. Teacher effects on student achievement have generally been looked at through three areas of research. With most controlling for student demographic information, research has centered on variations between classrooms in student achievement, associating specific teacher characteristics with student achievement, and associating teacher practices with student achievement (Konstantopoulos & Chung, 2011). Examining variations between classrooms assumes that systematic differences among classes in achievement are due to teacher effectiveness (Nye, Konstantopoulos, & Hedges, 2004). This approach exhibited evidence for differences of student achievement between classrooms but lacks the ability to say what teacher differences contribute to the findings.

Teacher characteristics that have been researched for their effects on student achievement include teacher experience, education, content knowledge, salary, and motivation, to name a few (Konstantopoulos, 2014). Studying low performing, high poverty schools, Huang & Moon (2009) found that the number of years teaching a specific grade had an influence on student achievement but licensing, education level, and total years teaching did not. Due to the type of experience being an important factor, the authors also looked at what could be driving the difference. The “seasoned” grade

level teachers spent more time in group level instruction than the less experienced grade level teachers, which may have had an impact on the results.

While there may be underlying mechanisms that contribute to the relationship between teacher salary and student achievement, there is evidence that the relationship is strong. A study on class size and teacher salary relating to student achievement showed that a teacher's salary was just as strongly associated with math and reading achievement as having smaller class sizes (Peevely, Hedges, & Nye, 2005). The authors suggest that more work be done to determine if salary is based on higher skill, educational attainment, experience, or a district's willingness to pay more and how those factors could lead to this effect.

Research concerning ethnic achievement gaps has shown that a teachers explicit and implicit prejudiced attitudes and expectations were associated with student achievement in math and reading (Peterson, Rubie-Davies, Osborne, & Sibley, 2016). The authors found that students performed better when the teacher's expectations were high and implicit biases favored their own ethnic group. However, a teacher's explicit biases were found to be unrelated to performance after controlling for previous performance.

Specific curricula and techniques such as scaffolding have been used to demonstrate how teacher practices can influence student achievement. Agodini and Harris (2016) researched four math curricula and found that even though the effectiveness of the curricula were moderated by teacher characteristics such as teacher knowledge, two of the programs were more effective across all of the conditions studied.

This finding highlights the importance of selecting a program that can be robust against differences between teachers and classrooms. Accompanying the idea that specific curricula can influence student achievement is the idea that support given during teaching can have important impacts. Scaffolding, or support designed to transfer responsibility of learning, has been shown to be effective depending on the type of support and the amount of independent working time for students (Pol, Volan, Oort, & Beishuizen, 2015). Pol et al. found that highly tailored support for students was beneficial when there was more independent working time, which meant less instruction. Less tailored support was beneficial when there was reduced independent working time, which meant that students received more frequent help. The authors suggested that this finding could be due to highly tailored support allowing a more thorough understanding, if given time to process the information.

While the research on teacher effects shows significant impacts on student achievement, Good (2014) has described how effects are not always stable. Good explains that, like many high performing athletes, effectiveness can change from year to year. Differences in effectiveness, as measured by student achievement scores, may be due to student factors outside of teacher control.

Student factors. Student factors thought to influence achievement include ethnicity, disability status, limited English proficiency, mobility, and socioeconomic status as evidenced by their inclusion in legislation and school accountability subgroup metrics. These student factors are also featured in the statistical models of interest in the current study. Research has supported the effects of student factors with one study

finding that 78% of the variance of student achievement was attributed to student characteristics among student, classroom, and teacher factors (Huang & Moon, 2009). The authors noted that this finding was in line with various other studies that partitioned out similar factors.

Another study involving three years of state-wide reading achievement data in Kentucky demonstrated how the vast majority of variability in their hierarchical linear model was accounted for within schools instead of between schools or districts (Adelson, Dickinson, & Cunningham, 2016). Specifically, it was found that a student's prior reading achievement accounted for most of the variability in all three levels (students, schools, and districts), followed by student demographics/characteristics, school characteristics, and district characteristics. The study also found that being male, Black, an English Language Learner, or qualifying for free/reduced priced lunch led to lower reading achievement scores, on average. This finding had the authors suggest that interventions be at the classroom or student level instead of at the school or district level as typical improvement efforts tend to focus on.

The following sections review recent research in the study of ethnicity, disability status, limited English proficiency, mobility, and socioeconomic status (SES) for their effects on student achievement.

Recent ethnicity and student achievement centered research has investigated gaps between ethnic groups and potential causes or factors that could account for such gaps. Curran and Kellogg (2016) conducted research on White student versus minority student achievement gaps in science for kindergarten and 1st graders. Significant gaps between

White students and Asian, Hispanic, and Black students were found that were larger than gaps shown in reading and math. When controlling for SES, gaps were still present between White and minority students but at a smaller amount.

Even when the focus is on already high achieving students, ethnic gaps persist and grow larger throughout high school. Kotok (2017) found that student course tracking, individual SES, school SES, and immigration status played roles in widening gaps between ethnic groups. Course tracking, or course placement, creates automatic limits to achievement levels due to the sequential nature of high school courses. The initial course placement of students can either allow opportunities for advanced placement exams and eventual knowledge growth or hinder them. High achieving Asian students were far more likely than White students to be enrolled in advanced courses and Asian and White students were even more likely than Black or Hispanic students to be enrolled in advanced courses. Individual and school level SES was found to be connected to ethnic achievement gaps, with lower SES being associated with lower achievement in both cases. Unfortunately, even when minority students were enrolled in higher SES schools, they tended to be placed in lower level courses than non-minority students. Finally, Asian advantages over White students may be due to immigrant status instead of a straight ethnic difference. It was found that higher achievement over White students disappeared when the Asian students were not immigrants.

At the kindergarten through 5th grade levels, mean achievement gaps exist between ethnicities and when comparing the highest and lowest performing within each ethnicity (Davis-Kean & Jager, 2014). A study exploring math and reading achievement

found that ethnic standings differed when comparing high and low achievers. In math for high performers, Asian students scored barely above White, Black, and Hispanic students who performed similar. In math for low performers, large gaps were seen between Asian students, followed by Black students, and then similar performing White and Hispanic students. For reading high performers, White students scored higher than Asian and Hispanic students while all scored higher than Black students. For reading low performers, White students performed better than Asian students, followed by Black students, and then Hispanic students. These finding stress that ethnic group differences should be looked at based on achievement levels in order to properly design interventions.

Research on students with disabilities and their achievement can differ by how studies group students. Studies can simply differentiate students by having or not having disabilities, use limited disability categories such as language impairment, or use multiple different disabilities categories (Stevens, Schulte, Elliott, Nese, & Tindal, 2015). Categories are wide-ranging and can include speech/language impairments, visual impairments, emotional disturbance, health impairments, intellectual disabilities, autism, learning disabilities, and hearing impairments.

In a study that examined achievement and growth gaps of 3rd through 7th graders, distinct differences between student with and without disabilities were seen. Schulte, Stevens, Elliott, Tindal, and Nese (2016) analyzed the mean achievement scores on a reading comprehension state test for gifted students, general education students, and eight types of students with disabilities groups. The study focused only on students who took

the regular forms of the reading comprehension assessment and not any alternative versions. The differences between achievement scores were stable across years for all groups including all the different disability groupings. The growth of each of the groups were similar as well, with students making larger gains in earlier years than later years. The results indicated that while there were small amounts of narrowing gaps between general education groups and students with disabilities groups, due to the initial achievement levels, no gaps were meaningfully closed.

Lower achievement levels were also seen in research that focused on the interaction between disability status and minority status in Kindergarten through 5th grade. The goal of the research was to determine if there were longitudinal and multiplicative effects of having both an IEP and being a minority student. No interaction was found for disability status and minority status, however findings indicated that disability status and minority status were independently related to lower math and reading achievement over time (Wu, Morgan, & Farkas, 2014). The authors noted that the effects of the factors on achievement may be additive but disability status with minority status was not disproportionately associated with lower achievement.

Limited English proficient (LEP) students, or English language learners (ELL), are students who speak a non-English language and are not proficient enough in English to perform classwork (Slama, 2012). Even though some students may not reach English proficiency during their school tenure, some states remove the label of LEP after a specified time limit (Cawthorn, 2010) or if the student meets a specified level of achievement on state mandated assessments (Rossell, 2006) which limits the

accommodations that are provided. While the conditions of being considered LEP differ among states, the subgroup is considered one of the fastest growing student populations in the U.S. (Young, Cho, Ling, Cline, Steinberg, & Stone, 2008).

Research focusing on growth patterns of achievement in math achievement have shown gaps between native English speakers and LEP students. Using longitudinal data on 3rd through 8th graders, LEP students had significantly lower starting levels but had similar growth rates over time (Ding & Davison, 2004). LEP students having growth rates to match native English speakers is encouraging but with the initial low achievement, it means that they will be unable to catch up.

Kieffer (2008) found similar results for Kindergarten through 5th graders in reading achievement. Students who were LEP had lower achievement levels and did not close the gap with English proficient speakers. Diving deeper into the association of demographic factors on achievement, the author looked at low SES and LEP status together. Findings suggest that LEP status moderated the effects of poverty on achievement by LEP students having less negative effects of poverty than native English speakers. However, the achievement levels of LEP students in high SES schools were consistently low so the results may be more due to being "underserved" in those schools.

Student mobility can be defined multiple ways depending on the focus of the investigation. Categories of mobility can include a change in residence, movement within the school system, movement outside the school system, change during the school year, change during the summer, promotional moves, non-promotional moves, and school-choice moves (Grigg, 2012; Parke & Kanyongo, 2012). Research on student mobility has

typically focused on non-promotional school changes and has been generally shown to have negative consequences for students such as lowered achievement or dropping out (Anderson, 2017).

Dauter and Fuller (2016) summarized research on the prevalence of student mobility. At the student level, research has demonstrated negative behavioral, academic, and dropout incidence effects for those with high mobility. At the community level, mobility was shown to be more common in low-incomes families, schools with low levels of material resources, and specific racial groups. Additionally, schools that have larger than 50% minority populations, urban schools, and high schools experience higher mobility rates than their counterparts (Han, 2014).

The effects of mobility on students are not only short-term but long-term. After controlling for prior achievement and behavior, mobile students were more likely to develop behavior problems, were less engaged in school, and have slower reading growth compared to non-mobile students (Lleras & McKillip, 2015). Mobility has also been linked to persistently lower achievement in addition to potential spillover effects on teachers and fellow students (Iserhagen & Bulkin, 2011). Through interviews of teachers and principals in high mobility schools, educators expressed the difficulties of "catching-up" mobile students. This catching-up process influenced materials teachers presented to their entire classrooms by needing to accommodate the mobile students.

While research has clearly demonstrated potential adverse outcomes due to mobility, at least one study has shown that early involvement strategies may help to lessen the potential negative outcomes. Herbers et al. (2012) demonstrated that the

negative achievement outcomes due to mobility could be mitigated by having higher early reading achievement. Early higher reading scores predicted higher reading achievement for all students but was more significant for mobile students.

While the previous factors have been fairly easy to identify if a student is part of a specific grouping, SES definitions can vary depending on the education system's or researcher's desired measures. Common definitions of SES include parent education, parent occupation/occupational prestige, household income, free/reduced price lunch status, or a composite of the previous (Dickinson & Adelson, 2014). Decades of research has shown that low-SES students are consistently outperformed by high-SES students in all subjects and grade areas (Taylor, 2005). The following is a selection of recent research that show the impact of SES on achievement within SES grouping, grade, and subject combinations.

A study examining the impact of SES on achievement within elementary, middle, and high school students in New Jersey found that SES accounted for a large amount of the variance in achievement (White et al., 2016). This effect was seen to increase as grade level increased indicating that SES can have a larger impact in middle and high school level students. Using the percent of students who qualified for free/reduced price lunch, the authors looked at the change in R^2 from a base model to a model with free/reduced price lunch percentage. The base model included percent female, faculty mobility, enrollment, and average class size. Large increases in accounted for variance were shown for each grade band ranging from .41 to .57 in reading and .36 to .53 in

math. The increases were so large and consistent that the authors suggested reevaluating the role of high stakes testing and the tests themselves.

Toutkoushian and Curtis (2005) examined multiple indicators of districts' SES levels for their influence on district level mean achievement test scores and postsecondary outcomes. The authors used district unemployment rate, percentage of adults with a bachelor degree or further, and percentage of students qualified for free or reduced-price meals as predictor variables. Findings included that over half of the variability in district achievement scores was accounted for by the combination of the three SES measures. Similarly, over half of the variability in predicting the percent of students taking the SAT and attending a 4-year college was accounted for. The study noted that the models lacked including any underlying factors that could be associated with SES such as parent involvement. Matching the previously reviewed article, the authors stated that it is "unfair" to lower SES districts to compare them to higher SES districts due to the conclusions found in this and much other research.

Through years of research, student factors have been established to contribute heavily to current and future student achievement. At the present time, a large number of states do not include factors such as socioeconomic status and mobility directly into accountability measurements and instead simply compare demographic groupings on the present measures and indicators. Potential models like the ones to be tested in the current study may provide a way to measure schools so that these types of factors are not forgotten in the high-stakes accountability environment.

Generalizability Theory

To determine variation sources between schools, models, and years, a generalizability study will be used. Generalizability studies, or G studies, are analyses used to estimate the reliability of measurements and their multiple sources of error (Alkharusi, 2012) based on Generalizability theory. Brennan (2011) described Generalizability theory (GT) as a continuation of Classical Test Theory (CTT) and ANOVA that focuses on variance components. A G study uses data, typically a sample, "from a universe of admissible observations that consists of facets" which are used to estimate the variance components (Brennan, 2003). G studies can have different forms but common arrangements are fully crossed one- or two- facet designs. Examples for a one-facet G study (*p x i*) would be where multiple individuals/persons (*p*) are given identical assessments (*i*). One-facet fully crossed G study designs are statistically identical to intraclass correlations and Cronbach's alpha to estimate within observer reliability (Mushquash & O-Conner, 2006). A two-facet G design (*p x i x r*) could be a setup of multiple individuals/persons (*p*) given identical sets of assessments (*i*) which are scored by multiple raters (*r*). In a two-facet fully crossed G study the observed score variance is made up of seven components that are estimated through random-effects ANOVA's expected mean squares:

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(r) + \sigma^2(pi) + \sigma^2(pr) + \sigma^2(ir) + \sigma^2(pir,e).$$

The variance values are used to determine the percentage of variance each component accounts for. The main effect $\sigma^2(p)$ component refers to variance attributed to differences in a person's scores, $\sigma^2(i)$ refers to difference between items and $\sigma^2(r)$ refers to difference

between raters. The two-way interaction component $\sigma^2(pi)$ refers to variations in items for a person averaging over raters, $\sigma^2(pr)$ is the differences in raters for a person averaging over items, and $\sigma^2(ir)$ is the differences in raters and items averaging over persons. Finally, the three-way interaction of $\sigma^2\ (pir,e)$ is the synonymous to random error (Lakes & Hoyt, 2009).

Once the variance components are estimated, a Decision study (D study) can be conducted. D studies are used to estimate a Generalizability Coefficient for norm-referenced/relative decisions or an Index of Dependability for criterion-referenced/absolute decisions, both of which are similar to the reliability coefficient in CTT (Webb & Shavelson, 2005). Changes in sample sized from G to D studies can allow researchers to test if increasing or decreasing facet counts will improve or lessen the variability of the object of measurement. Additionally, using the Satterthwaite approximation for degrees of freedom with G study mean square values, a minimum reliability standard for the study can be obtained (Gugiu & Gugiu, 2017). In order to calculate the standard a margin of error, or decision criterion, must be set by the researcher. The decision criterion set should reflect the scale being used as a researcher would not want to have a criterion that includes too large of a proportion of the scale or so small that it would be unduly difficult to obtain the minimum reliability standard. The minimum reliability standard can be compared with the Generalizability coefficient to determine if the D study has reached acceptable reliability, which may or may not be in line with the historical 0.70 or 0.80 rules of thumb.

Chapter 3. Methodology

Data Source

Data used in the analyses was obtained directly from the Ohio Department of Education's (ODE) publicly available data portal on ODE's website. The Advanced Data portal includes data elements such as enrollment, attendance, demographic information, financial data, test scores, and teacher data at the state, district and school levels. The data selected can be exported into files for further analysis. Unique district and school identification numbers are used to combine the data elements.

Data is collected from schools and districts through the Education Management Information System (EMIS). Depending on the information collected, data is reported monthly, quarterly, by semester, or yearly. Districts and schools must submit data according to ODE's standards and formats. Data collection includes demographic information, attendance, course information, financial data, and test results which are then blended by ODE. ODE is responsible for the publication of available data. Data validation processes are conducted through EMIS as an initial audit of the data for districts and schools so they may adjust any missing or conflicting information. However, the accuracy and completeness of data is the responsibility of the individual districts and schools.

There is a total of 3,866 individual schools that are reported on the ODE data portal. Not all of these schools are included in the analyses due to incomplete reporting by the districts and schools, and/or not having been rated on performance measures at the discretion of ODE. Examples of schools that are not a part of the analyses include Drop-

Out Recovery (DORP) schools, certain Special Education schools, or schools that have been closed in previous years but are still included in the reporting tables. Data publicly available from ODE is restricted when demographic groups include less than 10 students. Count data and their associated percentages are not shown for groupings with less than 10 students. Reported percentages show no higher than 95% for any given group, although percentages less than 5% are reported.

Student groupings are dictated by ODE and are defined by specific standards. A student is defined as being economically disadvantaged, or low socioeconomic status, in Ohio by either being eligible for free or reduced-price lunch, being in a household that has a recipient for public assistance programs, or whose guardians have completed a Title 1 application and meet the income guidelines for that year. Disability status is determined by being officially identified as deaf, blind, visually impaired, speech or language impaired, orthopedic impaired, emotionally disturbed, cognitively disabled, autistic, having a traumatic brain injury, developmentally delayed, possessing a specific learning disability, or having other health impairments. Limited English proficiency is determined by administration of an assessment that measures proficiency in English reading, writing, listening, and speaking.

Performance Index

The Performance Index is a sub-measure of the Achievement Component on Ohio's school report cards (Ohio Department of Education, 2015a). Every state mandated assessment required to be given to a student at a given school is classified into an Achievement Level. Each Achievement Level is associated with a point value. Point

values are detailed in Table 1. The point values are then multiplied by the percentage of assessment that fall within the Achievement Level. For example, if 80% of assessment scores are Proficient and 20% score Accelerated, then (80*1.0) + (20*1.1) = 80 + 22 = a Performance Index score of 102.

Achievement Level	Point Value
Advanced Plus	1.3
Advanced	1.2
Accelerated	1.1
Proficient	1.0
Basic	0.6
Limited	0.3
Untested	0.0

Table 1. Performance Index Levels

The minimum points possible is 0.0 and maximum number of points possible is 120.0. The minimum number of points would be achieved if all students in the school did not take any state mandated assessments. Students who are on a formal acceleration plan and take an assessment that is at a higher grade than enrolled in receive a higher Achievement Level, provided that the true level is at least Proficient. This allows an assessment to be considered Advanced Plus by the accelerated student scoring Advanced on the assessment. Although technically the maximum possible points would be 130.0, "a PI Score of 120 is considered to be a perfect score because this score would be earned if 100% of the tests from nonaccelerated students were into the Advanced range" (Ohio Department of Education, 2017).

Letter grades are given for the Performance Index scores based on the percent of points the school has received. An A if given for 90-100 percentage points, a B for 89-

89.9 percentage points, a C for 70-79.9 percentage points, and D for 50-69.9 percentage points, and an F for 0.0 to 49.9 percentage points.

Assessments given in the 2011-2012, 2012-2013, and 2013-2014 school years were the Ohio Achievement Assessments (OAAs) and the Ohio Graduation Tests (OGTs). The OAAs included reading and math for 3rd through 8th graders and science for 5th and 8th graders. The OGTs included reading, math, writing, science, and social studies for 10th graders. Only 10th graders or first time OGT takers are included in the Performance Index. Although only those student "count," those who do not pass the OGTs their first administration will retake the assessment until reaching proficient in order to meet graduation requirements.

Performance Index scores are reported on ODE's report card website for all applicable schools and districts. Only overall scores are available and there are no subgroup PI scores calculated.

Ohio Coalition for Quality Education Model

The OCQE model is a school-level multiple linear regression model that includes four predictor variables. All predictor variables and the dependent variable are averaged the latest three school years. To be included in the model a school must have three years of Performance Index (PI) scores and reported at least one year of the predictor variables. The model is as follows:

$$\textit{Performance Index Score} = b_1(ED) + b_2(SWD) + b_3(LEP) + b_4(MOB) + a$$

where *ED* is Economically Disadvantaged percentage, *SWD* is Students with Disability percentage, *LEP* is Limited English Proficiency percentage, and *MOB* is the percentage

of students who have been enrolled at least one year (mobility). All percentages are based on school enrollment and are reported at least annually by each school to the ODE.

To determine each school's grade based on the OCQE model the predicted PI score is compared to the observed PI score. Grades range from A-F, mirroring the current grading scale in place by ODE. An A is given for schools that have observed PI scores more than 10 points above their predicted PI, a B is given for schools that have observed PI scores between 5 and 10 points above their predicted PI, a C if given for schools that have observed PI scores between 5 points above and 5 points below their predicted scores, a D is given for schools with observed PI scores between 5 and 10 points below their predicted scores, and an F is given for schools with observed PI scores more than 10 points below their predicted scores.

For the present analyses, the OCQE model has been created using one year's data instead of averaged three years. This will allow for more meaningful and accurate comparisons between the OCQE model and the SSM.

Similar Students Measure

The SSM is a school-level weighted multiple linear regression that is separated by grade span and school year. The grade spans considered in the model are Elementary, Middle and High school which comprise of Kindergarten through 5th grade, 6th grade through 8th grade, and 9th grade through 12th grade, respectively. The model is run for each school year, given a performance category, then the performance category is aggregated amongst the latest three school years to fall within a performance band. To be included in the SSM a school must have a PI score, not serve an at-risk population such

as juvenile schools or a state authorized special education school, or not have fewer than 50 state-mandated tested students.

The following descriptions of variables included in the model are presented as defined by ODE and not necessarily defined as they are in California. Slight modifications have been implemented to fit the SSM model and will be discussed where applicable. The model is as follows;

$$\textit{Performance Index Score}^* = b_1(ED)^* + b_2(SWD)^* + b_3(LEP)^* + b_4(MOB)^* + b_5(AS)^* + b_6(BL)^* + b_7(HI)^* + b_8(MU)^* + b_9(WH)^* + a$$

where *ED* is Economically Disadvantaged percentage, *SWD* is Students with Disability percentage, *LEP* is Limited English Proficiency percentage, *MOB* is the percentage of students who have been enrolled at least one year (mobility), AS is percentage of students reporting as Asian, BL is percentage of students reporting as Black, HI is percentage of students reporting as Hispanic, MU is percentage of students reporting as multi-racial, WH is percentage of students reporting as white, and * denotes the weight for a given case. The regression is weighted by the total number of students tested in each school. All percentages are based off total required tested students and not general enrollment. Grade span is determined by the grade span in the school which has the most "required to test" students.

A necessary difference between the original SSM model and the presented model was that less ethnicity variables are included in the presented model. For the ethnicities of Pacific-Islander and American-Indian, only 4 schools reported any percentage of students so the variables are not included. Also, the original SSM included the actual number of tested students. Due to ODE's available data and reporting rules on counts and

percentages, the "total required to test" was used in its place. Additionally, the original SSM included a set of grade span models that included a parental education level variables. ODE does not have this data publicly available and/or does not collect this data so it was not included in the present analysis. The above differences from the original model would also have to be made if ODE were to implement the SSM model. This creates a "pseudo-SSM" since it deviates from the technical guidelines of the original model.

To determine the performance categories, an individual prediction interval at 68% (1 standard error) and 95% (two standard errors) is calculated for each school. Table 2 describes the performance categories and Table 3 describes the performance bands in detail.

Category	Description
Far Above	School's PI score was more than two standard errors above the predicted PI score
Above	School's PI score was between one and two standard errors above the predicted PI score
Within	School's PI score was within one standard errors above and below the predicted PI score
Below	School's PI score was between one and two standard errors below the predicted PI score
Far Below	School's PI score was more than two standard errors below the predicted PI score

Table 2. SSM Performance Categories

Band	Description
Far Above All Years	School in performance category of Far Above all three years
Above All Years	School in performance categories of Above or Far Above for all three years
Above Most Years	School in performance categories of Above or Far Above in two out of three years
Within/Fluctuating	School in performance categories of Within two out of the three years with no more than one year in a Below or Above category
Below Most Years	School in performance categories of Below or Far Below in only two out of three years
Below All Years	School in performance categories of Far Below and Below for all three years
Far Below All Years	Schools in performance category of Far Below for all three years

Table 3. SSM Performance Bands

Procedure

In order to properly compare the models, the OCQE model has been recreated using single year data, instead of the three-year average, and the SSM's one-year data will be used. Schools with all six data points from the OCQE model and SSM were included in the analyses. The models were analyzed by comparing model coefficients and fit. Correlations between each model's predicted Performance Index scores and observed Performance Index scores are also calculated. Subgroup comparisons including socioeconomic status levels, school type, and location typology are examined by model. Additionally, model assumptions are tested to determine if there are violations.

A generalizability study was conducted to determine sources of systematic variance. The generalizability study was a balanced, two-facet design with schools being the object of measurement and the facets being the models/measure (OCQE, SSM, and Performance Index) and academic year. The sources of variance of interest include the

variability due to schools, the models/measure, the years, the interactions of the facets, and unaccounted error. Schools, as the object of measurement, is always considered random. The facet of academic year is also considered random as the years measured could be expected to change. The facet of models/measure was considered fixed as they are not representative of other models and there is no attempt to generalize to other models.

The D study sample sizes were adjusted from the G study values. Year count was increased to 6 and model count was decreased to 1. Year count was increased to increase the variance around years and model count was decreased to eliminate variance around models. Selection of the minimum reliability standard decision criterion will be discussed in the results section. Follow-up Cronbach's alpha statistics were computed by model to test reliability differences of school scores between years.

Chapter 4. Results

Results for descriptive statistics, correlation analyses, regression analyses, and model validation were completed through use of SPSS 24.0. Generalizability study results were completed through use of SPSS 24.0 and the Generalizability Theory Workbook excel file that was created by Gugiu, Gugiu, and Baldus (2012).

For the OCQE model data, manual calculations were needed. Due to reporting errors of data percentages, Limited English Proficiency percentages were calculated from enrollment counts. Mobility percentages were calculated by combining the percentage of students who were enrolled 1 to 2 years and enrolled 3 or more years. Potential minimum and maximum percentages for Students with Disabilities enrollment and Economically Disadvantaged enrollment were 0.0% and 95.0%, respectively. Potential minimum and maximum percentages for Limited English Proficiency enrollment and Mobility enrollment were 0.0% and 100.0%, respectively.

For the SSM data, manual calculations were also needed. All factors were calculated using "required to test" counts by grade level as percentages reported were at the grade and subject level, and not the overall level. Ethnicity percentages are based on the reported data and may not total to 100% as some ethnicities were below the reportable count threshold (less than 10) or may have been omitted by the reporting school. Potential minimum and maximum percentages for all SSM factors range from 0.0% to 100.0%.

Additionally, each SSM year has three models within it separated by grade span. Many schools in Ohio do not fit the typical parameters of Kindergarten to 5th grade for

elementary, 6th through 8th grade for middle, and 9th through 12th grade for high schools. In order to determine the appropriate grade span of schools that served students from Kindergarten to 12th grade, 5th to 9th grade, or other mixture of grades, counts of students by individual grade level were used. For example, if in a 5th through 12th grade school most students were enrolled in 9th through 12th grade, then that school was categorized as a high school. Due to changing enrollments it is possible that a school is categorized as one grade span for a year and then categorized as another grade span in a subsequent year.

A school was included in the analyses if there were 3 years of both OCQE and SSM data for comparison purposes. Roughly 3,200 schools have a reported Performance Index per year and the present analyses included 3,012. Common reasons for not having the necessary data included newly formed charter schools, too few students to be included in SSM (under 50), or incomplete data reporting.

Descriptive Statistics

OCQE. As shown in Table 4, descriptive statistics for each factor were fairly stable throughout the three years of data for the OCQE model. The Performance Index scores ranged from 36.34 to 118.49 on a scale with a potential maximum of 120 and had standard deviations that ranged from 10.94 to 11.53. Students with Disabilities percentages were nearly identical throughout the three years with means from 14.9% to 15.4% and standard deviations from 7.9% to 8.1%. The Economically Disadvantaged factor had the most evenly distributed percentages amongst schools with means of 49.7% to 51.6% and standard deviations of 25.9% to 26.9%. The least evenly distributed

percentages were Limited English Proficient factors with means of 2.0% to 2.4% and standard deviations of 6.2% to 6.8%. Depending on the school year, 73% to 76% of schools reported having zero Limited English Proficient students. Mobility percentages had the smallest range with only 29.9% to 44.8% separating the minimum and maximum values.

Students with Disabilities, Limited English Proficient, and Mobility percentages were heavily skewed. Based off of the given skewness values, histograms of the factors, and Shapiro-Wilks tests (all $p<.001$), there was confirmation that the factors were not normally distributed. However, this effect may be due to outliers in the factors. Students with Disabilities factors displayed a small proportion of schools reporting percentages at the higher end of the range. Limited English Proficient factors had one notable outlier reporting 100%, with the next highest reported percent being 58%-66%, depending on the school year. The Mobility factors had 5 potential outliers that may have contributed to the skewness. The factors themselves are measures of special populations so it may not be surprising that there were substantial numbers of schools that had large/small populations of any given factor or that there would be certain schools that were likely to serve special populations.

Year	Variable	M	SD	Min.	Max.	Skew
2011-2012	Performance Index	95.91	10.94	48.66	118.49	-1.37
	Students with Disabilities	0.154	0.079	0.017	0.950	5.52
	Economically Disadvantaged	0.497	0.259	0.018	0.950	0.29
	Limited English Proficient	0.020	0.062	0.000	0.967	5.49
	Mobility	0.972	0.023	0.638	1.000	-4.16
2012-2013	Performance Index	95.49	11.31	37.41	118.20	-1.31
	Students with Disabilities	0.153	0.079	0.018	0.950	5.47
	Economically Disadvantaged	0.506	0.262	0.016	0.950	0.27
	Limited English Proficient	0.022	0.064	0.000	1.000	5.34
	Mobility	0.973	0.022	0.701	1.000	-4.08
2013-2014	Performance Index	95.66	11.53	36.34	117.00	-1.31
	Students with Disabilities	0.149	0.081	0.020	0.950	5.34
	Economically Disadvantaged	0.516	0.269	0.017	0.950	0.27
	Limited English Proficient	0.024	0.068	0.000	1.000	5.06
	Mobility	0.974	0.023	0.552	1.000	-5.44

**Note: N=3012*

Table 4. OCQE Model Descriptive Statistics

SSM. Descriptive statistics for the SSM's three grad span models are displayed in Table 5, Table 6, and Table 7 for total required to test students. The three grade spans saw similar statistics and outlier issues for the factors of Performance Index, Students with Disabilities, Economically Disadvantaged, Limited English Proficient, and Mobility. Students with Disabilities factors again had a small number of schools reporting higher percentages. Limited English Proficient factors had one to three major outliers for all

years and grade spans with the exclusion of 2011-2012 middle school and 2013-2014 elementary school models. Across years, 43%-46% of schools reported not having any Limited English Proficient required to test students. Mobility factors also had one potential outlier for the three years of high school and one for the 2011-2012 and 2012-2013 middle school years.

The required to test ethnicity percentages included in the SSM are Asian, Black, Hispanic, Multi-Racial, and White. Only Black and White percentages ranged from 0.0% to 100.0%. Hispanic percentages went to a maximum of 90.5% in elementary school but only 45.7% in middle school, and 49.2% in high school. All other ethnicities had maximums below 43.0%. Skewness values were high for most grade span and years of ethnicity due a large portion of schools reporting 0.0% for ethnicities. Overall, approximately 86% of schools reported no required to test students for the ethnicities of Asian, 53% for Black, 68% for Hispanic, 55% for Multi-Racial, and just 5% for White.

Total Required to Test counts had means of 240.7 to 245.0 for elementary school, 500.6 to 504.4 for middle school, and 214.9 to 229.5 for high school. Standard deviations ranged from 141.6 to 221.9 for elementary school, 250.8 to 256.2 for middle school, and 214.9 to 229.5 for high school. One charter school in particular had a Total Required to Test count of 7,051 for elementary school in 2011-2012, switched span to high school with 7,925 in 2012-2013, and then back to elementary school in 2013-2014 with 7,254. Two other schools had counts much higher than most other schools. One school categorized as elementary had counts between 5,118 and 6,300, with the other being in

high school that ranged between 2,597 to 2,672. The next few closest Total Required to Test schools' counts were approximately 1,500 to 1,800 students.

Grade Span categorization was mainly stable for the schools included in the analyses. Out of the 3,012 schools, 1,666 were in elementary, 588 were in middle school, and 693 were in high school. The 65 schools that changed categories within the three years switched between two categories. No school fluctuated between all three categories. Thirty-six of the 65 schools were 5th through 6th grade schools. The remaining fluctuating schools were a blend of 15 other grade span ranges such as Kindergarten through 12th grade, 5th through 8th grade, and 4th through 7th grade.

Year	Variable	M	SD	Min.	Max.	Skew
2011 - 2012	Performance Index	94.21	12.07	48.66	118.49	-1.12
	Students with Disabilities	0.162	0.084	0.024	1.000	4.485
	Economically Disadvantaged	0.560	0.278	0.018	1.000	0.063
	Limited English Proficient	0.032	0.074	0.000	0.920	4.911
	Mobility	0.955	0.033	0.719	1.000	-1.629
	Asian	0.008	0.028	0.000	0.295	4.906
	Black	0.195	0.300	0.000	1.000	1.519
	Hispanic	0.033	0.080	0.000	0.873	4.082
	Multi-Racial	0.036	0.047	0.000	0.273	1.317
	White	0.687	0.323	0.000	1.000	-1.033
	Total Required to Test	245.01	218.49	52.00	7051.0	19.14
2012 - 2013	Performance Index	93.64	12.39	37.41	118.20	-1.07
	Students with Disabilities	0.160	0.077	0.030	1.000	3.983
	Economically Disadvantaged	0.566	0.281	0.000	1.000	0.062
	Limited English Proficient	0.033	0.071	0.000	0.665	3.857
	Mobility	0.957	0.031	0.712	1.000	-1.637
	Asian	0.008	0.029	0.000	0.339	4.945
	Black	0.194	0.298	0.000	1.000	1.523
	Hispanic	0.038	0.086	0.000	0.889	3.886
	Multi-Racial	0.036	0.046	0.000	0.264	1.240
	White	0.682	0.322	0.000	1.000	-1.004
	Total Required to Test	240.70	141.57	51.00	2608.0	4.27
2013 - 2014	Performance Index	93.97	12.70	36.34	117.00	-1.06
	Students with Disabilities	0.161	0.083	0.010	1.000	4.282
	Economically Disadvantaged	0.575	0.288	0.016	1.000	0.057
	Limited English Proficient	0.036	0.075	0.000	0.661	3.675
	Mobility	0.958	0.032	0.758	1.000	-1.634
	Asian	0.009	0.033	0.000	0.389	4.861
	Black	0.194	0.296	0.000	1.000	1.517
	Hispanic	0.041	0.088	0.000	0.905	3.604
	Multi-Racial	0.037	0.048	0.000	0.435	1.490
	White	0.675	0.322	0.000	1.000	-0.980
	Total Required to Test	243.28	221.93	54.00	7254.0	19.68

**Note: 2011-2012: n=1694, 2012-2013: n=1689, 2013-2014: n=1699*

Table 5. SSM Model Descriptive Statistics, Elementary Schools

Year	Variable	M	SD	Min.	Max.	Skew
2011 - 2012	Performance Index	97.31	8.28	62.42	112.08	-1.52
	Students with Disabilities	0.154	0.090	0.050	1.000	6.511
	Economically Disadvantaged	0.443	0.214	0.018	1.000	0.366
	Limited English Proficient	0.016	0.035	0.000	0.328	4.212
	Mobility	0.962	0.031	0.523	1.000	-5.827
	Asian	0.010	0.025	0.000	0.217	3.729
	Black	0.101	0.207	0.000	1.000	2.739
	Hispanic	0.024	0.047	0.000	0.415	4.068
	Multi-Racial	0.033	0.034	0.000	0.200	1.178
	White	0.809	0.230	0.000	1.000	-2.024
	Total Required to Test	504.42	250.81	55.00	1792.0	0.90
2012 - 2013	Performance Index	96.55	8.74	57.85	111.71	-1.60
	Students with Disabilities	0.157	0.102	0.045	1.000	5.856
	Economically Disadvantaged	0.459	0.228	0.017	1.000	0.475
	Limited English Proficient	0.017	0.044	0.000	0.705	8.313
	Mobility	0.964	0.027	0.558	1.000	-6.022
	Asian	0.011	0.027	0.000	0.241	3.908
	Black	0.103	0.207	0.000	1.000	2.688
	Hispanic	0.027	0.051	0.000	0.406	3.807
	Multi-Racial	0.033	0.035	0.000	0.200	1.265
	White	0.804	0.234	0.000	1.000	-1.955
	Total Required to Test	503.80	254.86	58.00	1827.0	0.95
2013 - 2014	Performance Index	97.00	8.96	57.03	112.99	-1.58
	Students with Disabilities	0.154	0.088	0.032	0.983	6.062
	Economically Disadvantaged	0.468	0.240	0.018	1.000	0.569
	Limited English Proficient	0.017	0.042	0.000	0.579	6.319
	Mobility	0.965	0.026	0.777	1.000	-2.296
	Asian	0.011	0.028	0.000	0.284	4.689
	Black	0.102	0.206	0.000	0.990	2.719
	Hispanic	0.029	0.055	0.000	0.457	3.888
	Multi-Racial	0.034	0.035	0.000	0.181	1.142
	White	0.801	0.236	0.000	1.000	-1.940
	Total Required to Test	500.59	256.16	60.00	1815.0	0.99

**Note: 2011-2012: n=618, 2012-2013: n=622, 2013-2014: n=619*

Table 6. SSM Model Descriptive Statistics, Middle Schools

Year	Variable	M	SD	Min.	Max.	Skew
2011 - 2012	Performance Index	98.81	9.20	52.87	113.14	-1.76
	Students with Disabilities	0.154	0.075	0.019	0.867	3.951
	Economically Disadvantaged	0.425	0.231	0.000	1.000	0.698
	Limited English Proficient	0.014	0.039	0.000	0.595	7.653
	Mobility	0.961	0.046	0.487	1.000	-4.271
	Asian	0.005	0.021	0.000	0.273	6.424
	Black	0.127	0.255	0.000	0.995	2.158
	Hispanic	0.017	0.051	0.000	0.492	4.759
	Multi-Racial	0.015	0.031	0.000	0.211	2.522
	White	0.795	0.276	0.000	1.000	-1.750
	Total Required to Test	214.93	236.76	52.00	5118.0	13.52
2012 - 2013	Performance Index	99.02	9.50	50.35	115.12	-1.63
	Students with Disabilities	0.151	0.084	0.010	0.983	5.071
	Economically Disadvantaged	0.437	0.241	0.015	1.000	0.724
	Limited English Proficient	0.012	0.039	0.000	0.661	9.154
	Mobility	0.963	0.045	0.513	1.000	-4.048
	Asian	0.006	0.025	0.000	0.338	8.469
	Black	0.126	0.254	0.000	1.000	2.197
	Hispanic	0.018	0.051	0.000	0.481	4.258
	Multi-Racial	0.016	0.032	0.000	0.224	2.313
	White	0.792	0.276	0.000	1.000	-1.743
	Total Required to Test	229.46	395.03	52.00	7925.0	15.42
2013 - 2014	Performance Index	98.61	9.68	44.14	114.47	-1.70
	Students with Disabilities	0.148	0.085	0.005	0.903	3.796
	Economically Disadvantaged	0.452	0.252	0.018	1.000	0.714
	Limited English Proficient	0.013	0.038	0.000	0.461	6.848
	Mobility	0.965	0.039	0.578	1.000	-3.630
	Asian	0.006	0.029	0.000	0.471	9.430
	Black	0.128	0.255	0.000	1.000	2.177
	Hispanic	0.021	0.055	0.000	0.479	4.286
	Multi-Racial	0.018	0.033	0.000	0.194	2.060
	White	0.786	0.279	0.000	1.000	-1.702
	Total Required to Test	227.31	277.61	51.00	6300.0	15.62

**Note: 2011-2012: n=700, 2012-2013: n=701, 2013-2014: n=694*

Table 7. SSM Model Descriptive Statistics, High Schools

Research Question 1

To answer Research Question 1, *to what degree do school-level demographic characteristics influence school achievement scores in Ohio and how much do demographic characteristics account for variability in the OCQE model and SSM,* regression analyses were conducted. The OCQE and SSM model were recreated as specified above and the results are detailed in Table 8, Table 9, Table 10, and Table 11 below.

OCQE. The OCQE multiple regression models had high R^2 values with 0.71 for 2011-2012 ($F(4,3007)=1871.56$, $p<.001$), 0.70 for 2012-2013 ($F(4,3007)=17.86.94$, $p<.001$), and 0.71 for 2013-2014 ($F(4,3007)=1881.79$, $p<.001$). These findings show that a very large proportion of the variability in achievement scores are accounted for by the included demographic factors.

All coefficients for all years were statistically significant at the $p<.001$ level, with the exception of the 2011-2012 Limited English Proficient factor which was significant at the $p<.01$ level. Directionally, Students with Disabilities, Economically Disadvantaged, and Limited English Proficient factors had negative coefficients and Mobility had positive coefficients. It would be expected based on the literature that having students with disabilities, low incomes, and limited English would decrease achievement scores. Mobility findings also match the literature that suggests students who stay with a school year over year would perform better on achievement assessments. Mobility was positive in the models due to the language in the model specifications for the SSM and OCQE models. The specifications stated that mobility is the amount retained instead of the

students who were mobile. This is in opposition to the other factors where the "at-risk" group is the counted percentage.

Based on the Standardized Beta coefficients, Economically Disadvantaged factors was the strongest factor associated with Performance Index score with Beta coefficients of -0.655, -0.651, and -0.621 for each successive year. The least associated factor was Limited English Proficient with -0.030, -0.063, and -0.065 for each successive year.

Year	Variable	B	SE B	β	R^2
2011-2012	Constant	42.184	5.795		0.71
	Students with Disabilities	-24.902	1.505	-0.181***	
	Economically Disadvantaged	-27.704	0.510	-0.655***	
	Limited English Proficient	-5.312	1.795	-0.030**	
	Mobility	73.500	5.779	0.153***	
2012-2013	Constant	45.607	6.199		0.70
	Students with Disabilities	-27.068	1.576	-0.190***	
	Economically Disadvantaged	-28.110	0.531	-0.651***	
	Limited English Proficient	-11.058	1.799	-0.063***	
	Mobility	70.354	6.197	0.134***	
2013-2014	Constant	25.418	6.072		0.72
	Students with Disabilities	-26.496	1.585	-0.185***	
	Economically Disadvantaged	-26.691	0.520	-0.621***	
	Limited English Proficient	-11.099	1.712	-0.065***	
	Mobility	90.571	6.048	0.178***	

**Note: p<0.05, **p<0.01, ***p<.001*

Table 8. OCQE Model Multiple Regression Results

SSM. The SSM weighted multiple regression models displayed large R^2 values that increased with the grade span. The elementary school models had R^2 values of 0.79 to 0.80, middle school models had values of 0.81 to 0.83, and high school models had

values of 0.84 to 0.86. Model *F*-ratios for all years and grade spans were significant at the $p<.001$ level.

2011-2012 School Year Models. For the elementary school model, all coefficients were statistically significant not including Limited English Proficient and Asian factors. All ethnicity factors had negative coefficients which is unexpected, especially for the White coefficient. The Standardized Beta coefficients revealed the strongest factors to be Black (β = -0.763), Economically Disadvantaged (β = -0.429), and White (β = -0.423). The weakest significant factor was shown to be Multi-Racial (β = -0.046).

In contrast to the elementary school model, the middle school and high school models had only Students with Disabilities, Economically Disadvantaged, and Mobility as statistically significant factors. Each model had Economically Disadvantaged (middle β = -0.540, high β = -0.510) as the strongest factor followed by Students with Disabilities (middle β = -0.173, high β = -0.271), then Mobility (middle β = 0.147, high β = 0.241).

2012-2013 School Year Models. All model coefficients were statistically significant for the elementary school model with the exception of Limited English Proficient and Multi-Racial. Again, all ethnicity factors had negative coefficients and the strongest factors were Black (β = -0.744), Economically Disadvantaged (β = -0.465), and White (β = -0.407). The weakest significant factor was Asian (β = -0.039). The middle school model mirrored the 2011-2012 middle school model with the only significant factors being Economically Disadvantaged (β = -0.479) followed by Students with Disabilities (β = -0.230), then Mobility (β = -0.111).

The high school model had all factors being statistically significant except for Asian and Multi-Racial. Similar to the elementary school model, the strongest factors were Black (β = -0.784), White (β = -0.632), and Economically Disadvantaged (β = -0.472). The weakest significant factor was Limited English Proficient (β = -0.089).

2013-2014 School Year Models. The elementary school model was comparable to the 2012-2013 elementary school model with non-significant factors being Limited English Proficient and Multi-Racial. Also similar were the strongest and weakest significant factors being Black (β = -0.793), White (β = -0.478), Economically Disadvantaged (β = -0.452), and Asian (β = -0.045). Middle school, once again, had three significant factors which included Economically Disadvantaged (β = -0.453), Mobility (β = 0.244), and Students with Disabilities (β = -0.201).

For the high school model, Asian, Hispanic, and Multi-Racial were non-significant factors. The strongest factors were Black (β = -0.482) and Economically Disadvantaged (β = -0.465). The weakest significant factor was Limited English Proficient (β = -0.125).

Summary. All SSM models displayed very large R^2 values. While there were differences between the significance and importance of factors between the grade span models, they were mostly similar across years. The elementary school models consistently displayed the highest number of significant factors and middle school models displayed the least. Across all year and grade span models, Economically Disadvantaged was among the strongest predictors. Mobility coefficients were positively oriented. Students with Disabilities and Economically Disadvantaged factors continually

had negative coefficients. When significant, Limited English Proficient and Ethnicity factors had negative coefficients as well.

Grade Span	Variable	B	SE B	β	$R^2(F)$
Elementary	Constant	87.014	6.260		0.79 (689.4***)
	Students with Disabilities	-26.573	2.242	-0.155***	
	Economically Disadvantaged	-17.782	0.845	-0.429***	
	Limited English Proficient	2.237	2.642	0.013	
	Mobility	41.301	4.162	0.128***	
	Asian	-9.303	6.011	-0.026	
	Black	-30.566	4.710	-0.763***	
	Hispanic	-26.400	4.845	-0.174***	
	Multi-Racial	-12.292	5.035	-0.046*	
	White	-15.901	4.705	-0.423***	
Middle	Constant	57.324	12.071		0.83 (326.5***)
	Students with Disabilities	-26.571	3.158	-0.173***	
	Economically Disadvantaged	-18.929	1.057	-0.540***	
	Limited English Proficient	1.455	4.832	0.007	
	Mobility	46.076	7.742	0.147***	
	Asian	17.591	10.025	0.069	
	Black	0.524	9.417	0.013	
	Hispanic	4.962	9.867	0.028	
	Multi-Racial	11.115	9.035	0.046	
	White	9.154	9.465	0.259	
High	Constant	91.126	6.033		0.84 (408.09***)
	Students with Disabilities	-39.678	2.692	-0.271***	
	Economically Disadvantaged	-21.072	1.073	-0.510***	
	Limited English Proficient	-5.801	5.401	-0.023	
	Mobility	23.068	1.642	0.241***	
	Asian	12.547	8.031	0.034	
	Black	-4.731	5.812	-0.120	
	Hispanic	-5.580	6.342	-0.030	
	Multi-Racial	9.723	6.349	0.034	
	White	1.371	5.777	0.037	

**Note: p<0.05, **p<0.01, ***p<.001*

Table 9. SSM Model Weighted Multiple Regression Results, 2011-2012 School Year

Grade Span	Variable	B	SE B	β	R^2(*F*)
Elementary	Constant	100.183	7.195		0.79 (702.7***)
	Students with Disabilities	-29.428	2.403	-0.159***	
	Economically Disadvantaged	-19.415	0.814	-0.465***	
	Limited English Proficient	-2.164	2.786	-0.012	
	Mobility	28.065	5.576	0.068***	
	Asian	-13.644	6.009	-0.039*	
	Black	-30.773	4.817	-0.744***	
	Hispanic	-24.062	4.877	-0.165***	
	Multi-Racial	-5.808	4.990	-0.021	
	White	-15.677	4.820	-0.407**	
Middle	Constant	75.927	13.071		0.81 (284.6***)
	Students with Disabilities	-33.402	3.082	-0.230***	
	Economically Disadvantaged	-16.628	1.002	-0.479***	
	Limited English Proficient	-6.989	4.919	-0.034	
	Mobility	40.612	8.463	0.111***	
	Asian	15.971	11.172	0.065	
	Black	-16.002	10.653	-0.384	
	Hispanic	-12.104	11.107	-0.071	
	Multi-Racial	3.424	9.901	0.014	
	White	-4.817	10.705	-0.132	
High	Constant	111.067	6.056		0.86 (456.8***)
	Students with Disabilities	-37.136	2.734	-0.245***	
	Economically Disadvantaged	-19.407	1.021	-0.472***	
	Limited English Proficient	-22.097	5.515	-0.089***	
	Mobility	27.655	1.692	0.285***	
	Asian	-14.256	7.857	-0.040	
	Black	-33.189	5.723	-0.784***	
	Hispanic	-25.973	6.099	-0.129***	
	Multi-Racial	4.027	6.435	0.013	
	White	-24.927	5.696	-0.632***	

**Note: p<0.05, **p<0.01, ***p<.001*

Table 10. SSM Model Weighted Multiple Regression Results, 2012-2013 School Year

Grade Span	Variable	B	SE B	β	$R^2(F)$
Elementary	Constant	76.903	6.509		0.80 (728.0***)
	Students with Disabilities	-25.301	2.267	-0.140***	
	Economically Disadvantaged	-18.578	0.766	-0.452***	
	Limited English Proficient	-2.234	2.717	-0.013	
	Mobility	54.749	4.474	0.154***	
	Asian	-14.160	5.615	-0.045*	
	Black	-33.850	4.656	-0.793***	
	Hispanic	-26.392	4.660	-0.183***	
	Multi-Racial	-8.669	4.810	-0.032	
	White	-18.700	4.659	-0.478***	
Middle	Constant	26.959	14.078		0.82 (300.3***)
	Students with Disabilities	-31.548	3.338	-0.201***	
	Economically Disadvantaged	-15.578	0.997	-0.453***	
	Limited English Proficient	-5.022	5.034	-0.025	
	Mobility	84.162	8.708	0.244***	
	Asian	13.843	11.033	0.057	
	Black	-6.585	10.681	-0.152	
	Hispanic	-3.117	10.927	-0.019	
	Multi-Racial	13.490	9.800	0.053	
	White	1.649	10.687	0.044	
High	Constant	96.669	6.576		0.84 (385.3***)
	Students with Disabilities	-38.456	2.808	-0.265***	
	Economically Disadvantaged	-18.551	1.041	-0.465***	
	Limited English Proficient	-31.876	6.819	-0.125***	
	Mobility	29.994	2.209	0.245***	
	Asian	-1.476	7.843	-0.005	
	Black	-20.144	6.251	-0.482**	
	Hispanic	-10.579	6.614	-0.057	
	Multi-Racial	7.709	6.888	0.026	
	White	-12.473	6.239	-0.323*	

**Note: p<0.05, **p<0.01, ***p<.001*

Table 11. SSM Model Weighted Multiple Regression Results, 2013-2014 School Year

Research Question 2

To answer research question 2, *what are the relationships between the OCQE model, SSM, and the Performance Index rating*, correlations and descriptive statistics were ran for the outcomes of each model. In addition, the subgroups of school SES level, school type, and location typology were examined.

To determine SES level, school Economically Disadvantaged percentiles were assessed. The percentiles were fairly close, plus or minus 2 percentage points, between years so an approximation of the 25th and 75th percentiles were used as dividing points. The 2013-2014 Economically Disadvantaged data was used to categorize schools. High SES is considered 0.0%-29.9% Economically Disadvantaged, Medium SES is considered 30.0%-69.9%, and Low SES is considered 70.0%-100.0%. There were 727 (24.1%) schools in the High SES group, 1496 (49.7%) in the Medium SES, and 789 (26.2%) in the Low SES group.

School type refers to traditional versus charter schools. The Ohio Department of Education (2016) defines charters schools as:

> Community schools, often called charter schools in other states, are public nonprofit, nonsectarian schools that operate independently of any school district but under a contract with an authorized sponsoring entity that is established by statute or approved by the State Board of Education. Community schools are public schools of choice and are state and federally funded.

In the present study there are 155 charter schools, 85% of which serve multiple grade spans.

Location typology refers to ODE's classification of districts', and therefore the schools they service, demographic and geographic attributes (Ohio Department of Education, 2015b). The attributes of interest comprise of measures such as poverty level, population density, median income, and property values, among others. The data for each of the attributes is collected from sources such as the Census Bureau, the Ohio Department of Taxation, and the Ohio Department of Education. The last update of typology was conducted in 2013. The major groupings of typology are Rural, Small Town, Suburban, and Urban. Nearly all charter schools have not been given a typology although the majority of charter schools are located in urban areas. In the present study, 671 (22.3%) schools are categorized as Rural, 712 (23.6%) as Small Town, 764 (25.4%) as Suburban, 711(23.6%) as Urban, and 154 (5.1%) are not categorized. The methodology for categorizing districts is available on the Ohio Department of Education's website. Due to the uncategorized location typology group comprising of 99% charter schools, any findings from the subgroup will not be discussed.

Correlations. Correlations between Performance Index scores, the OCQE model predictions, and the SSM model predictions were all positive, high, and significant at the $p<.001$ level. For comparison purposes, correlations between the three years of Performance Index scores were correlated from $r = 0.946$ to $r = 0.965$. The lowest correlation of any subgroup's yearly Performance Index scores was Rural at $r = 0.773$, although the majority were in the high 0.8's to low 0.9's. Looking at all schools, the OCQE predicted scores were correlated between years at $r = 0.959$ to $r = 0.972$. The SSM predicted scores were correlated between years at $r = 0.965$ to 0.974. The OCQE

predicted scores were correlated with Performance Index scores from r = 0.833 to r = 0.845 and the SSM predicted scores were correlated with Performance Index scores from r = 0.873 to r = 0.891. The OCQE predicted scores were correlated with SSM predicted scores from r = 0.913 to r = 0.948.

For subgroup comparisons, same year correlations will be examined. For example, OCQE 2011-2012 versus SSM 2011-2012 correlations will be of interest and not OCQE 2011-2012 versus SSM 2012-2013. Traditional schools had correlations that expectedly matched the strength of the overall findings since they comprise of 95% of the schools analyzed. Charter schools had lower, but still strong correlations between scores. The OCQE and SSM predicted scores were correlated from r = 0.819 for 2013-2014 to r = 0.868 for 2011-2012. Correlations between both model's predicted scores and Performance Index scores dropped nearly 0.2 compared to traditional schools. Values ranged from r = 0.624 to r = 0.653 for the OCQE predicted scores and r = 0.661 to r = 0.703 for SSM.

For the correlations between the OCQE and SSM predicted scores, Low SES ranged between r = 0.680 for 2013-2014 to r = 0.799 for 2011-2012, Medium SES ranged between r = 0.841 for 2013-2014 to r = 0.861 for 2011-2012, and High SES ranged between r = 0.798 for 2013-2014 to r = 0.872 for 2011-2012. Performance Index scores were more correlated with the SSM model predictions than for the OCQE model predictions. Performance Index and SSM had values from r = 0.610 for Medium SES to r = 0.740 for Low SES while the OCQE values were from r = 0.542 for Low SES to r = 0.631, also for Low SES.

Location typology groups shared similarities in correlations between scores. Rural, Small Town and Suburban had OCQE and SSM correlations above r = 0.90 and Urban ranged from r = 0.849 to r = 0.889. Correlations for Performance Index and OCQE scores ranged from r = 0.695 to r = 0.781 for all but Rural whose values were between r = 0.536 to r = 0.556. Correlations for Performance Index and SSM scores were more mixed with Rural having the lowest values of r = 0.536 to r = 0.556, Small Town between r = 0.715 to r = 0.746, Suburban between r = 0.802 to r = 0.808 and Urban between r = 0.817 to r = 0.827.

Overall, correlations of the scoring methods were very strong and stable between years. While a large portion of values were above r = 0.8, subgroups had values around the 0.5's and 0.6's that displayed that there is variability between the observed and predicted scores for specific populations. Interestingly, for charter and SES subgroups there were decreasing correlations at higher years. Location typologies did not have as defined of a trend and were mostly mixed in terms of increasing or decreasing among the school years.

Score Descriptive Statistics. The observed Performance Index scores for the given years ranged from 48.66 to 118.49. The predicted scores for the OCQE model ranged from 27.38 to 113.73 and for the SSM from 50.65 to 113.31. Table 12 describes the means and standard deviations for the models and subgroups. Within groupings the means and standard deviations are very similar, most within 2-3 points. However, between groupings there are noticeable differences. For the SES categories, Low SES had the lowest means, followed by Medium SES then High SES. Charter schools had the

lowest means out of all subgroups for nearly all measures and years. The typology subgroup means has clear distinctions with Urban being the lowest followed by Rural, Small Town, then Suburban. For all subgroups, higher means were associated with lower standard deviations indicating that the higher scores, or predicted scores, were accompanied by less variability between schools.

Model	Subgroup									
	All Schools		Low SES		Medium SES		High SES		Charter	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PI 2011-2012	95.91	10.94	82.51	11.17	98.52	5.16	105.09	3.53	81.80	11.93
PI 2012-2013	95.49	11.31	81.43	11.12	98.28	5.40	105.02	3.83	81.33	11.15
PI 2013-2014	95.66	11.53	81.26	11.20	98.52	5.53	105.42	3.90	80.65	12.10
OCQE 2011-2012	95.91	9.24	83.96	5.60	97.09	4.41	106.48	3.15	83.69	9.26
OCQE 2012-2013	95.49	9.49	82.97	5.38	96.79	4.42	106.42	2.93	82.70	9.45
OCQE 2013-2014	95.66	9.75	82.49	5.44	97.21	4.29	106.78	2.61	82.60	9.51
SSM 2011-2012	95.80	9.72	82.92	7.60	97.92	4.32	105.40	2.96	81.35	9.05
SSM 2012-2013	95.34	10.10	81.67	7.65	97.73	4.33	105.24	2.87	79.44	9.66
SSM 2013-2014	95.59	10.26	81.50	7.36	98.09	4.34	105.72	2.70	80.21	9.50

Model	Subgroup							
	Rural		Small Town		Suburban		Urban	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PI 2011-2012	98.63	4.90	99.64	5.07	103.51	5.11	84.49	11.97
PI 2012-2013	98.33	5.15	99.42	5.32	103.36	5.44	83.45	12.35
PI 2013-2014	98.66	5.27	99.50	5.53	103.74	5.59	83.50	12.37
OCQE 2011-2012	97.52	5.08	98.25	6.21	103.77	5.70	86.22	6.84
OCQE 2012-2013	97.08	5.43	97.93	6.34	103.46	5.96	85.72	7.07
OCQE 2013-2014	97.40	5.69	98.44	6.44	103.60	5.88	85.51	7.54
SSM 2011-2012	98.68	3.94	99.28	4.98	102.94	5.62	85.00	8.57
SSM 2012-2013	98.31	4.11	99.27	5.01	102.53	5.81	84.29	8.92
SSM 2013-2014	98.61	4.53	99.59	5.11	102.89	5.83	84.18	9.04

Table 12. Descriptive Statistics for Subgroups by Model and Year

The differences between the predicted scores of the OCQE model and SSM are displayed in Table 13. These differences were calculated by subtracting the predicted SSM score from the OCQE predicted score. Positive means would indicate that the OCQE model rated schools higher and negative means would indicate that the SSM rated schools higher.

Group	Model Year					
	2012		2013		2014	
	Mean	Range	Mean	Range	Mean	Range
All Schools	0.1179	29.52	0.1550	27.23	0.0738	60.32
Low SES	1.0366	29.52	1.3005	27.02	0.9873	56.96
Medium SES	-0.8337	19.69	-0.9485	22.66	-0.8868	28.95
High SES	1.0792	13.89	1.1826	12.23	1.0588	21.75
Rural	-1.1614	14.01	-1.2305	14.98	-1.2125	16.69
Small Town	-1.0300	19.54	-1.3364	15.02	-1.1513	17.32
Suburban	0.8321	19.69	0.9379	18.36	0.7051	22.45
Urban	1.2174	21.58	1.4308	22.32	1.3284	56.96
Charter	2.3408	29.52	3.2586	27.23	2.3982	41.37

Table 13. Difference of Scores between OCQE and SSM by Year

Most predicted scores were relatively similar between models based on means that were mostly within two points, but there were a handful of inconsistencies. Large ranges, for example in 2014 with ranges as high as 60.32, were due to differences in the demographic enrollment and demographic required to test counts. In addition, only two schools had a difference over 20 points between the models and both were in 2014 which contributed to larger ranges for the All Schools, Low SES, Urban, and Charter school categories for that year. While still higher, removing the two outliers would provide ranges closer to the previous years.

In summary, examining the descriptive statistics of the observed and predicted scoring indicates how similar the models perform. Most of the differences appear to be between subgroups instead of between models. Subgroups that have been shown to perform lower, such as Urban or Low SES populations, were predicted to perform at similar rates with the two proposed models. However, the differences of predicted scores showcases how the models can operate differently for those schools who have special populations that may not contribute to student achievement indices by not being required to test.

Research Question 3

To answer research question 3, *are the OCQE model and SSM adequate representations of the given data,* regression model assumptions were tested. Multiple regression analyses have four main assumptions consisting of linearity, homoscedasticity of residuals, independence of residuals, and normality of residuals (Cohen, Cohen, West, & Aiken, 2003). Linearity is tested through examination of scatterplots of factors against the dependent variable and residual values against predicted values. Homoscedasticity, or constant variance of residuals, is tested through examination of scatterplots of residual values against each factor and residuals against predicted values. Independence of residuals is tested through scatterplots of residuals against case numbers, and the Durbin-Watson statistic. Normality of residuals is tested through histograms and p-p plots of residuals. Finally, Variance Inflation Factors (VIF) will be analyzed for multicollinearity in the models.

OCQE. The three years of OCQE models were nearly identical when it comes to model assumptions so findings will be discussed at the model, not the year, level. The assumption of linearity was questionably met. The factors of Economically Disadvantaged and Mobility were linear but had large numbers of data points that were at higher percentages of the factors. Limited English Proficient factors did not have any distinct form of relationship, mainly due to most data points at 0% for the factor. Students with Disabilities factors were linear for the majority of data points but a very small, but separate, group clustered at the higher percentages of the factor which tilted the form of the relationship to be nonlinear. All-in-all, there did not seem to be an issue with linearity as much as there is significant clustering at high or low percentages of the factors.

Homoscedasticity was not met for Economically Disadvantaged factors with wide spreading of data points at the higher percentage levels. The three other factors did not show clear violations of the assumption, however again, most data points clustered at the higher or lower ends of the factors. Examining the residual and predicted values scatterplot, there was a distinct violation of the assumption. The scatterplot showed a sharp downward spread at predicted values of roughly 80 indicating heteroscedasticity and possible issues with linearity.

Independence of residuals was met by showing no significant pattern in the residual scatterplot and Durbin-Watson values at acceptable levels between 1.93 and 1.95. The normality of residuals was also met with a normal histogram of residuals and p-p plots where observed residuals followed the trend line. Multicollinearity did not seem to be an issue with VIF values between 1.07 and 1.55.

SSM. The SSM also had nearly identical results between years and even between grade span models. Findings will be discussed overall and any grade span differences will be covered when applicable.

The SSM model factors tended to be linear with clustering at extreme values or be ball-like clusters with no real relationship against the Performance Index. Economically Disadvantaged, Mobility, Black, and White factors displayed linearity but with bunching at extreme values. Limited English Proficient, Asian, Hispanic, and Multi-racial factors all had clusters with the majority of values at 0%. Students with Disabilities factors again had two separate groupings with most of the data points in a linear fashion but a small cluster that would alter the linearity of the factor.

Homoscedasticity fared better with the SSM than the OCQE model but did not meet the assumption. The scatterplot with residuals and predicted values still showed a spreading at lower predicted values but the spreading was more consistent without a sharp downturn. The Elementary grade span model was more distributed across predicted values while the Middle and High models were more clustered across predicted values. Similar to linearity findings, residual values against the factors were either linear or clustered. Economically Disadvantaged, Black, and White factors were mainly linear. Economically Disadvantaged factors had data point spreading at higher percentages but Black factors had many data points sitting at higher percentages and White factors had many data points at lower percentages. The rest of the factors were clustered with the majority of the data points falling at or near 0%.

The independence of residuals was adequate when looking at residuals against case numbers and the Durbin-Watson statistics. The Durbin-Watson values stayed between 1.78 and 2.15 for all models and years. The normality of residuals was also met as an assumption with normally distributed histograms of residuals and approximately straight observed residuals against the trend line in the p-p plots.

Unfortunately, there were major issues with multicollinearity based off of VIF values for the Black and White factors. The VIF values for Black factors ranged from 98.14 to 108.99 for Elementary, 198.89 to 207.9 for Middle, and 87.69 to 94.11 for High school grade levels. White VIF values ranged from 116.71 to 125.43 for Elementary, 254.33 to 274.46 for Middle, and 100.08 to 108.37 for High school grade levels. The extremely large VIF values are not completely surprising due to Black and White percentages in SSM being negatively correlated in the high 0.90's. Hispanic also was troublesome when it comes to VIF with the Middle grade span model having values from 10.65 to 15.27.

The OCQE model appears to have potential issues with the linearity of the factors and constant variance of residuals but not the independence and normality of residuals. The SSM fared the same with issues of heteroscedasticity, potential issues with linearity, and positive findings for independence and normality of residuals. The incredibly large VIF values demonstrated problems of multicollinearity for the ethnicity factors that would need to be addressed.

Research Question 4

To answer research question 4, *utilizing a generalizability study for the OCQE model, SSM model, and the Performance Index, which model is more reliable for making decisions based on school performance,* a generalizability study was conducted and model/measure specific reliability analyses were run. Due to the large number of schools, stratified random sampling without replacement was applied. The strata were based on grade span so that the same proportion of Elementary, Middle, and High schools were represented in the sample. Using the 2013-2014 grade span data, Elementary schools made up 56% of the population, Middle schools made up 21% and High schools made up 23%. Five samples of 150 schools were selected that combined to total 750 schools (25% of the 3,012 schools).

The G study was a two-facet, fully crossed design with schools (p) as the object of measurement and the facets being models (α) and years (β). Schools and year were considered random whereas models were considered fixed. A D study was conducted to determine if increasing the number of years would increase variability accounted for by model variations and if the study would meet minimum reliability requirements. Results for the G and D studies are displayed in Table 14.

The G study found that the majority of variance was accounted for by school (p) differences at 87.29%. There were essentially no mean differences among models (α) (0.00%) and only 0.05% attributed to mean differences across years (β). Variations for the school and model interaction ($p\alpha$) accounted for 9% which indicates slight differences in models scores for a school over years. Said differently, this would suggest that

schools' scores can differ between years within a model. Almost 1% of variance was accounted for by the school and year interaction (pβ) that shows scores minimal differences in years between models. Random error was low at 2.67% of total variance for the G study.

The D study year count was increased 6 to test if between model variations would increase. Additionally, model count was reduced to 1 because otherwise the analysis would be averaging across models which would be undesirable as each model is unique and there is not the need to generalize based on models. There was a slight increase for schools (*p*). and model (*pA*) and decrease in random error. Otherwise, all other variance sources remained similar. Based on the G study variance, results showed that there was not enough variability in years so an increase in D study count for years did not improve school variability to a high degree.

To determine a minimum reliability standard for the D study, a decision criterion of 10 was used. This margin of error was selected as roughly 90% of predicted scores were with +/- 5 points of the observed Performance Index score. Using the Satterthwaite approximation for degrees of freedom and the set decision criterion, the minimum reliability standard was found to be 0.923. The G coefficient for the D study was calculated to be 0.901 so the minimum reliability standard was not met, although it would be considered high by conventional norms.

G Study Source	SS	DF	MS	GVAR	%
School (*p*)	135477.79	149	909.25	97.00	87.29%
Model (α)	18.39	2	9.20	0.00	0.00%
Year (β)	65.87	2	32.94	0.06	0.05%
School x Model ($p\alpha$)	9840.21	298	33.02	10.02	9.01%
School x Year ($p\beta$)	1840.96	298	6.18	1.07	0.96%
Model x Year ($\alpha\beta$)	19.04	4	4.76	0.01	0.01%
Error ($p\alpha\beta$)	1769.60	596	2.97	2.97	2.67%
Total	149031.86	1349	110.48	111.13	100%

D Study Source*	DVAR	%
School (*p*)	96.85	90.05%
Model (*A*)	0.00	0.00%
Year (*B*)	0.01	0.01%
School x Model (*pA*)	3.34	9.31%
School x Year (*pB*)	0.18	0.17%
Model x Year (*AB*)	0.00	0.00%
Error (*pAB*)	0.16	0.46%
Total	100.69	100%

**Note: N (A) was decreased to 1 and for (B) was increased to 6*

Table 14. Two-facet Generalizability Analysis Results

To see if there were differences in reliability across years for each of the models, Cronbach's alpha was calculated. The alpha values were very similar between models for the three years with PI at 0.985, OCQE at 0.989, and SSM at .990. The very high, yet practically identical alpha's make it impossible to determine if one model is more reliable.

The generalizability and decision studies indicated that the majority of variation is between schools and that across years and models there was high consistency. Based on the findings of this analysis, the models are unified and consistent on average. It would

be expected that models would vary between one another due to difference in the model factors and the SSM being weighted. However, due to generalizability and Cronbach's alpha focusing in mean scores, results indicated little variation. This finding is consistent with the descriptive statistics and correlational analyses presented earlier in the results chapter.

Chapter 5. Discussion & Conclusion

The present study's goal was to determine the utility of two models that are being investigated by the Ohio Department of Education for inclusion as a state accountability measure. The procedure of the study included examining the outcomes of the models, subgroup differences, suitability of the models, and sources of systematic error. Data was collected from the publicly available school-level data on the Ohio Department of Education's report card website. The data included the current achievement measure of Performance Index, demographic enrollment data, student required to test data, and school classification data. The models were recreated with Ohio school data and analyzed through descriptive statistics, correlation analyses, assumption testing, and generalizability studies.

Results from the present study are interpreted in the following section. Overall conclusions based on the study findings are discussed in terms of statistical and practical importance. In addition, study limitations and implication for future research are considered.

Discussion

The present study sought to investigate the utility of two proposed achievement accountability models. The recreation of the Ohio Coalition for Quality Education model and Similar Students Measure found that much of the variance in Ohio's Performance Index achievement measure was accounted for by demographic enrollment variables. The R^2s of the models ranged from 0.70 to 0.86, which are very large and not normally seen values. The large values indicated that teacher and school influences may not have as

much to do with achievement in Ohio as policy makers would suggest. These findings confirm previously reviewed research from White et al. (2016), Toutkoushian and Curtis (2005), and Huang & Moon (2009) that the majority of variance in assessment scores can be accounted for by demographic factors, most notably socioeconomic status.

In the present study, Economically Disadvantaged percentage was the strongest predictor in all model years of the OCQE model. All demographic factors included in the OCQE model were significant predictors with Economically Disadvantaged, Students with Disabilities, and Limited English Proficient having negative coefficients and Mobility having positive coefficients across years. The findings show that having larger percentages of special populations decrease achievement scores. This matches results from studies that showed students have lower achievement if they were considered limited English proficient (Kieffer, 2008 & Ding & Davison, 2004), mobile (Iserhagen & Bulkin, 2011), have a disability (Wu, Morgan, & Farkas, 2014 & Schulte et al., 2016), or are economically disadvantaged (Taylor, 2005).

It is important to note that only Economically Disadvantaged percentages had a normal distribution of scores. The other three factors, and the SSM Ethnicity factors, had the majority of percentages clustered around 0% or 100%. From a real-world standpoint this is expected because schools would be unlikely to have large percentages of their students being non-English proficient or having a disability. This most likely led to the results that, while most influential, the Economically Disadvantaged coefficients had B weights of around -27 so a 10% change is Economically Disadvantaged percentage only changed predicted PI by less than 3 points. In contrast to the Mobility B weights of 70-

90, depending on the year, which would make a 10% change in mobility equal to a 7- to 9-point change in predicted PI. So, while the factor Economically Disadvantaged had the most statistical strength, mobility contributed to larger changes in the predicted score of a given school.

The SSM was similar to the OCQE model for the shared demographic variables with the exception of Limited English Proficient significance. Only the 2012-2013 and 2013-2014 high school models had significant coefficients for the Limited English Proficient factors. Economically Disadvantaged was the strongest predictors for all years and grade span models. The factors of Economically Disadvantaged, Students with Disabilities, and significant Limited English Proficient had negative coefficients. The Mobility factors had positive coefficients. These findings further supported the previously review research on demographic effects on student achievement.

The SSM included ethnicity factors in the model which were not consistently a good fit and should be considered for removal. The ethnicity factors varied in significance between grade span models and years. Middle school models did not have any ethnicity factors prove to be significant but Elementary school models consistently had Black, Hispanic, and White factors with negative, significant coefficients. Although all significant ethnicity coefficients were negative, the White coefficients had B weights half of the size of Black and Hispanic coefficients. This would indicate that higher percentages of White students decreased PI scores less than that of higher Hispanic or Black student percentages.

Additionally, the ethnicity factors have built-in multicollinearity because the percentages are dependent on each other. The ethnicity factors for a school should sum to 100% since it is a total breakdown of all the students in the school by specified ethnic group. Most significantly, Black and White factor percentages were almost perfectly, negatively correlated and, consequently, created VIF values that were very extreme. Having variables that are dictated by one another is a design flaw and either removal or categorization into a minority or non-minority percentage is advisable. By creating one ethnicity based variable such as a total minority percentage, the model would still capture the data intended and create larger distributions for the one variable versus the five that were included. Having simple a total minority percentage may also be more palatable for policy makers since it does not have the affect of pinpointing specific ethnicity groups as larger or smaller detriments to achievement scores.

The OCQE, SSM, and PI scores were highly correlated with each other and, on average, had only small differences between them. Subgroup correlations showed less strength than overall correlations, but still high. There were notable subgroup variations showing Low SES schools performing lower than High SES schools across all models and years, on average. The present study also showed clear achievement differences between rural and urban schools versus suburban schools. Haifeng and Cowen (2009) described how research is clear that urban schools tend to perform lower than suburban schools but mixed on rural versus non-rural achievement. Studies that explored location based differences tended to show influences on achievement were due to minority population and poverty rather than location. Ohio's rural populations tend to be lower in

the socioeconomic status spectrum so the lower achievement found for rural schools could be due poverty based characteristics.

Deviations between PI and the proposed models occurred when schools had larger percentages of multiple special populations. Essentially, a model's predicted scorings would give a boost to schools with large amounts of multiple special populations. Since literature shows that students who are economically disadvantages, have a disability, are limited English proficient, and/or change schools are highly associated with lower academic performance, these models would put schools with large amounts of multiple special populations more inline accountability-wise with other schools. However, based on the current findings, it is far more difficult for schools to have meaningful increases in the predicted scores by having only moderate amounts of a special population or large membership of only one grouping.

Any differences between the outcomes of the OCQE model and SSM stem from the student data used in their calculations. The OCQE uses school population enrollment data whereas SSM uses required to test student data and is weighted by enrollment. The differences could be more influential for schools that have higher needs students with disabilities or limited English proficient students that get alternate assessments or waivers to testing. High schools could be more impacted by the data differences since, at the time of the presented data, only 10th graders were required to test. This would mean that the OCQE data would include 9^{th} through 12^{th} grade enrollment percentages but SSM data would only include 10^{th} grade students required to test. Also, because the SSM is weighted, more influence is given to larger schools. It would make more sense to use the

SSM methodology of model factors based on required to test students since that is population PI is calculated on. The PI calculation only includes those students who are required to test as each student is given a proficiency level/point value based on an assessment score and those who do not test are still counted but given a 0-point value. Therefore, it would be sound to use demographic percentages based off of the same set of students rather than overall enrollment data.

Unfortunately, the appropriateness of the OCQE model and SSM methodology is questionable. Consistent indications that heteroscedasticity was present for all models and years, along with non-linearity for many of the factors would suggest that remedies for these violations be attempted. Noticeably, many of the factors in the models had large amounts of schools with roughly 0% or 100% of students in a factor. This led to clustering of the data points and make it difficult to assess model estimate accuracy. There was also less variation and lower percentages of the factors, particularly percentage Economically Disadvantaged, at higher scores of PI. This suggested that schools who scored higher on PI had fewer, if any, students in the special populations tested.

The reliability of the models was tested using a generalizability study and Cronbach's alpha. The results of both of the analyses reinforced the previous findings that minimal differences occurred between models and years tested. Nearly all of the variation when considering schools, models, years, and their interactions was accounted for by school differences. In earlier analyses, models showed very little differences between each other and only slightly more between years. The main differences that were found were not due to model differences but were between subgroupings of schools such

as suburban versus urban or high SES versus low SES. This was reflected in the results of the G study by 87% of variation coming from schools. The next largest accounted for variance was the interaction between schools and models with 9% which shows how schools scores can diverge between years within the models. Changes in student cohorts or student achievement between grades could explain the interaction variation. Random error, or unmeasured facets of measurement, only accounted for under 3%, followed by the minimal variation between schools and years with 1%.

The G study and other analyses in the present research imply that even though the tested models include demographic data as their predictors, they are not exempt of demographical differences in their outcomes. Again, because the variability within most of the factors is low for the majority of schools, en masse the influence of those factors is weak. There were relatively few schools that had large percentages of students in the special population categories. The most powerful factor, Economically Disadvantaged, was normally distributed so that most schools had moderate amounts of students who fell into the category. In other words, the high variability of Economically Disadvantaged percentages coupled with low variability in other factor percentages resulted in predicted scores that were more influenced by Economically Disadvantaged. However, this also created predicted scores that did not deviate significantly from their observed scores. Therefore, when looking at the model predicted scores on average, little differences emerge so deviations within or between the models are masked. Additionally, the OCQE models, SSMs, and PI scores were very consistent across years as shown by Cronbach's

alpha values. The consistency was so high that one model did not stand out as a more reliable option compared to the others.

Even though the demographic factors in the present study accounted for a large majority of variance in achievement scores, based on the reliability analyses and residuals of the models, there may be other factors that would contribute to achievement. Past research has shown how teacher characteristics such as teacher knowledge (Agodini & Harris, 2016) and instructional support (Pol, Volan, Oort, & Beishuizen, 2015) or school based characteristics such as class sizes (Peevely, Hedges, & Nye, 2005) and "School Choice" enrollment patterns (Ahn & McEachin, 2017) can affect achievement. The inclusion of additional demographic factors such as other measures of socioeconomic status or migrant status may increase the explanatory power of the models. While the current study is based mainly on Free and Reduced Price Lunch (FRPL) percentage as the definition of Economically Disadvantaged, the previous research discussed used varying measures of SES. Creating a composite measure of SES or using a different definition could bolster the results of the study and contribute to understanding the difference facets of SES that affect student achievement. Also, the original methodology of the SSM includes the creation of models with parental education as a factor. Parental education has been shown to be a consistent predictor of students' educational attainment in addition to the combination of parental aspirations and parental education being linked to students' performance (Spera, Wentzel, & Matto, 2007). Including a parental education factor in the OCQE model or SSM could contribute to understanding the influence of parents on student achievement.

Conclusion

The present study examined two demographically based models and Ohio's current achievement measure through regression analysis, correlational analyses, model/measure comparisons, subgroup differences, and reliability analyses to establish the validity of the models. Three years of data from 3,012 schools were used in the analysis. Demographic enrollment and required to test data for economically disadvantaged students, students with disabilities, limited English proficient students, mobility data, and ethnicity groups at the school level were used as predictors for the models.

Results indicated that model outcomes and significance were similar to each other and to the current achievement measure. The models and measures were highly correlated but lesser strength correlations were found within specific subgroupings. Across years, consistent predicted scoring and subgroup differences were found. Although the models showed statistical significance and consistency, the suitability of the models is in question due to assumption violations of heteroscedasticity, linearity, and multicollinearity. The factors in the models are measures of special populations which leads to clustered percentages, with most schools having a large or small population of any given group. The occurrence of clustering can lead to skewed and non-linear variables. Reliability analyses of a generalizability study and Cronbach's alpha illustrated the consistency of the models across years but was unable to designate a more reliable model due to model outcome similarities.

Overall, the proposed models aligned with previous research as to the effects of demographic variables on achievement. However, when examining subgroup differences, both models did not provide outcomes that heavily fluctuated from the current achievement measure. The reliability of the models was found to be satisfactory but the appropriateness of the use of multiple regression without transformations is in question. Further research is needed to correct assumption violations.

Gándara and Randall (2015) advised that accountability systems must have quality measures that are valid, reliable, transparent, and have positive outcomes that outweigh unintended or indirect negative consequences. Massive amounts of research have shown how tightly tied demographic characteristics are to measures of academic achievement. By understanding how various accountability measures assess schools and students, educational policies can better reflect true performance.

Limitations

A limitation of the present study is the selection of the schools and the school data used. Roughly 3,200 schools have an Performance Index score for a given year. Nearly 200 school were removed from the analysis due to incomplete data and the need to have three years' worth of data for comparison purposes. There may be systematic reasons why data was missing or not reported that could have influenced the outcomes of this study. Data imputation was not used in an effort to be consistent with the methodology that ODE would have practiced if the models were in use for the report card. Additionally, a few values for data points were extreme and may be due to reporting errors. One school in particular had a very large Students with Disabilities percentage that

could have been a reporting error. This data point led to extreme scores in the analysis and contributed to creating a larger range in predicted scoring. This school, in later years than in the study, was moved to an "ungraded" status and no longer receives the Performance Index.

Data included in the study was for the 2011-2012, 2012-2013, and 2013-2014 school years. During those years the Ohio Achievement Assessments (OAAs) were given for 3rd through 8th graders and the Ohio Graduation Tests (OGTs) were given for 10th graders. Beginning in the 2014-2015 school year, new assessments called the Ohio State Tests (OSTs) were given. These tests included English Language Arts and Mathematics for 3rd through 8th graders and Science for 5th and 8th graders. There are currently ten assessments that may be taken at specified time points throughout high school. The tests include Algebra I, Geometry, Integrated Mathematics I, Integrated Mathematics II, English Language Art I, English Language Arts II, American Government, American History, Biology, and Physical Science. The tests taken by high school students is dependent on the courses they are enrolled in throughout their tenure. The Performance Index will include data from all available assessments and now include data from all years of high school which is unlike the previous years. Additionally, the new high school assessment would influence the required to test counts and percentages used for the SSM. Due to the different assessments and extensive testing list for high schoolers, there may be differences in findings if the present study were to be rerun with the new assessment data.

The data available from ODE is only at the school level and not the student level. The influences of demographic characteristics on achievement may differ if we could analyze at the student level. There also could be compounding effects of demographics when a student is a member of multiple demographic groups such as being economically disadvantaged and having a disability. Additionally, there could be underlying factors that influence achievement that were not included in the model such as parent involvement or emphasis on schooling (Toutkoushian & Curtis, 2005).

The goal of the study was to replicate the proposed models in the method specified in House Bill 2 and California's methodology. However, multiple regression may be inappropriate for the data available. Issues with linearity/clustering, independence of residuals, and multicollinearity would lend to transforming the data. However, because the sample size was large, the sampling distribution is likely close to the true distribution so transformations may or may not assist in fixing assumption violations. Another issue is that schools are directly related to the districts they belong to. Schools within the districts likely share characteristics with each other more so than schools outside of the district. Information may be lost that would influence predicted score outcomes, relationships to achievement, and may make the models more robust. In order to determine just how related school scores were within districts, 4 randomly selected schools within the 247 districts with at least 4 schools were analyzed. Intraclass correlations (ICC) were conducted and found to be over 0.80 for all three models/measures used in the current study. This finding illustrates that schools within districts are highly similar when it comes to observed or predicted scoring of

achievement. Therefore, it would be strongly recommended that hierarchical linear modeling with schools nested within districts be attempted for the tested models if either were to be integrated into school evaluations.

Implications for Future Research

The present study focused on individual year data but the school outcomes would be based on averaged three-year data for OCQE or three-year combined data for the SSM, both of which would be categorized. Future research could analyze the category outcomes based on the multiple year data. Due to the findings that model outcome differences are negligible, research into the categorization of the model data may showcase schools that would gain or lose compared to current achievement measures. Additionally, the combination of three years of data may smooth out any outlying enrollment data so future research could determine single vs three-year differences.

Future research could examine the differences between the stated methodologies of the models and adjustments a version of the models that would better fit the data. Like mentioned in the limitations section, data transformations or hierarchical linear modeling could be attempted to see if there are ways to better describe the demographic data that is included in the models.

Analysis into how the proposed models interact with other accountability measures such as value-added or post-secondary outcomes could be of interest to policy makers, should the models be considered for inclusion in a state educational report card. Ohio's report card is extensive with many measures to rate schools upon. The differences between report card measures would have an impact on how to classify high or low

performing schools. Currently, sixteen states have begun to use A-F letter grades based on accountability measures to create an overall metrics on which to grade schools (Murray & Howe, 2017). Determining the viability of future models and the ability for them to create a cohesive picture of school performance would be important.

Further studies could be done on other accountability measures that would compare the current methodologies to a modified version that incorporated demographic data into the calculations. At the present time in Ohio and many other states, special demographic populations are only looked at as subgroups due to regulations the first came as part of the No Child Left Behind Act. The inclusion of demographic data into the calculations may reveal how much the measures are influenced by having larger amounts of subgroup populations such as students with disabilities. Specifically, the inclusion of a mobility variable would show the importance of changing school environments on achievement. In the present study and other research that was outlined in the literature review, the effects of mobility in achievement were significant and showed that students are negatively affected by changing schools. Mobility is a unique variable in that it is not a fixed characteristic like a disability and therefore would not be taken into account with previous achievement scores or would it be likely to be reoccurring. Achievement measures and growth measures such a Value-Added in Ohio could increase the accuracy of ratings and identify the magnitude of temporary events on achievement by including mobility in their calculations.

Further research could investigate students who were untested or who change schools after poor assessment scores. Due to the fact that Performance Index, and hence

the models, are dependent on the students who test, it would be prudent to look for potential trends into students who opt-out of testing or who move to different schools. Schools could potentially suggest that poor performing students opt-out or switch schools so the lesser scores do not influence the school's achievement results. This could also lead to the schools where the students end up receiving lower achievement scores. It is notable that charter schools perform lower than traditional schools. A part of the low performance could be due to the already poor performing students leaving their home school so the charters need to help students catch-up.

On a practical level, qualitative research into how parents and teachers are able to comprehend advanced calculations could be investigated. The purpose of school and district reports cards are, in part, a way to assist parents and educational professionals in understanding the successes and failures of school systems. It is important to keep measures clear, concise, and as transparent as possible so that individuals can make informed decisions. It could be of interest to study how parents and educational professionals synthesize and use the information given by accountability measures.

References

A-F Report Card – Ohio Department of Education. (2015a). Retrieved from http://education.ohio.gov/getattachment/Topics/Data/Accountability-Resources/A-F-Report-Card-2015-final.pdf.aspx

Adelson, J. L., Dickinson, E. R., & Cunningham, B. C. (2016). A multigrade, multiyear statewide examination of reading achievement: Examining variability between districts, schools, and students. *Educational Researcher, 45*(4), 258-262. doi:10.3102/0013189X16649960

Agodini, R., & Harris, B. (2016). How Teacher and Classroom Characteristics Moderate the Effects of Four Elementary Math Curricula.*Elementary School Journal*, *117*(2), 216-236.

Ahn, J., & McEachin, A. (2017). Student Enrollment Patterns and Achievement in Ohio's Online Charter Schools. *Educational Researcher*, *46*(1), 44-57.

Alexander, N. A., Sung, T. J., & Kankane, S. (2017). The Performance Cycle: The Association between Student Achievement and State Policies Tying Together Teacher Performance, Student Achievement, and Accountability. *American Journal Of Education, 123*(3), 413-446.

Alkharusi, H. (2012). Generalizability Theory: An Analysis of Variance Approach to Measurement Problems in Educational Assessment. *Journal of Studies in Education*, *2*(1), 184-196. doi:10.5296/jse.v2i1.1227

Anderman, E. M., Anderman, L. H., Yough, M. S., & Gimbert, B. G. (2010). Value-Added Models of Assessment: Implications for Motivation and Accountability. *Educational Psychologist, 45*(2), 123-137.

Anderson, S. (2017). School mobility among middle school students: When and for whom does it matter? *Psychology In The Schools,* doi:10.1002/pits.22010

Beating the Odds – Michigan Department of Education. (2013). Retrieved from http://www.michigan.gov/mde/0,1607,7-140-22709---,00.html

Brennan, R. L. (2003). Coefficients and Indices in Generalizability Theory. *CASMA Research Report, 1,* 1-44.

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education, 24,* 1-21. doi: 10.1080/08957347.2011.532417

Castellano, K. E., & Ho, A. D. (2013). A Practitioner's Guide to Growth Models. Council of Chief State School Officers.

Cawthon, S. W. (2010). Assessment Accommodations for English Language Learners: The Case of Former-LEPs. *Practical Assessment, Research & Evaluation*, 151-159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York, NY: Routledge.

Community School FAQ – Ohio Department of Education (2016). Retrieved from http://education.ohio.gov/Topics/Community-Schools/FAQs

Conley, D. C. (2015). A New Era for Educational Assessment. *Education Policy Analysis Archives, 23*(7-11), 1-40.

Curran, F. C., & Kellogg, A. T. (2016). Understanding Science Achievement Gaps by Race/Ethnicity and Gender in Kindergarten and First Grade. *Educational Researcher, 45*(5), 273-282. doi:10.3102/0013189X16656611

Dauter, L., & Fuller, B. (2016). Student Movement in Social Context. *American Educational Research Journal, 53*(1), 33-70.

Davis-Kean, P. E., & Jager, J. (2014). Trajectories of Achievement Within Race/Ethnicity: "Catching Up" in Achievement Across Time. *Journal Of Educational Research, 107*(3), 197-208.

Derthick, M., & Dunn, J. M. (2009). False Premises: The Accountability Fetish In Education. *Harvard Journal Of Law & Public Policy, 32*(3), 1015-1034.

Dickinson, E. E., & Adelson, J. J. (2014). Exploring the Limitations of Measures of Students' Socioeconomic Status (SES). *Practical Assessment, Research & Evaluation, 19*(1-4), 1-14.

Ding, C. S., & Davison, M. L. (2005). A longitudinal study of math achievement gains for initially low achieving students. *Contemporary Educational Psychology, 30*(1), 81-95. doi:10.1016/j.cedpsych.2004.06.002

Gándara, F. M., & Randall, J. J. (2015). Investigating the Relationship between School Level Accountability Practices and Science Achievement. *Education Policy Analysis Archives, 23*(112/113), 1-22.

Gietz, C., & McIntosh, K. (2014). Relations between student perceptions of their school environment and academic achievement. *Canadian Journal Of School Psychology, 29*(3), 161-176. doi:10.1177/0829573514540415

Goddard, Y., Goddard, R., & Kim, M. (2015). School instructional climate and student achievement: An examination of group norms for differentiated instruction. *American Journal Of Education, 122*(1), 111-131. doi:10.1086/683293

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). Growth model comparison study Practical implications of alternative models for evaluating school performance. Washington, DC: Council of Chief State School Officers.

Good, T. L. (2014). What Do We Know About How Teachers Influence Student Performance on Standardized Tests: And Why Do We Know So Little About Other Student Outcomes? *Teachers College Record, 116* (010303).

Grigg, J. (2012). School Enrollment Changes and Student Achievement Growth: A Case Study in Educational Disruption and Continuity. *Sociology Of Education, 85*(4), 388-404. doi:10.1177/0038040712441374

Gugiu, C. & Gugiu, M. (2017). Determining the Minimum Reliability Standard Based on a Decision Criteria. *Journal of Experimental Education.* doi.org/10.1080/00220973.2017.1315712

Gugiu, M. R., Gugiu P. C., & Baldus, R. (2012). Utilizing Generalizability Theory to Investigate the Reliability of Grades Assigned to Undergraduate Research Papers. *Journal of MultiDisciplinary Evaluation, 8*(19), 26-40. ISSN 1556-8180

Haifeng, Z., & Cowen, D. J. (2009). Mapping Academic Achievement and Public School Choice Under the No Child Left Behind Legislation. *Southeastern Geographer, 49*(1), 24-40.

Han, S. (2014). School Mobility and Students' Academic and Behavioral Outcomes. *International Journal of Education Policy & Leadership, 9*(6).

Hanley, A., Roehrig, A. D., & Canto, A. I. (2015). States' Expressed Versus Assessed Education Goals in the Era of Accountability: Implications for Positive Education. *Educational Forum, 79*(2), 130-147.

Herbers, J. E., Cutuli, J. J., Supkoff, L. M., Heistad, D., Chan, C., Hinz, E., & Masten, A. S. (2012). Early Reading Skills and Academic Achievement Trajectories of Students Facing Poverty, Homelessness, and High Residential Mobility. *Educational Researcher, 41*(9), 366-374

Huang, F., & Moon, T. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment, Evaluation & Accountability, 21*(3), 209-234. doi:10.1007/s11092-009-9074-2

Isernhagen, J. C., & Bulkin, N. (2011). The Impact of Mobility on Student Performance and Teacher Practice. *Journal Of At-Risk Issues, 16*(1), 17-24.

Itkonen, T., & Jahnukainen, M. (2007). An Analysis of Accountability Policies in Finland and the United States. *International Journal Of Disability, Development & Education, 54*(1), 5-23. doi:10.1080/10349120601149664

Jaekyung, L. (2007). Do National and State Assessments Converge for Educational Accountability? A Meta-Analytic Synthesis of Multiple Measures in Maine and Kentucky. *Applied Measurement in Education*, *20*(2), 171-203. doi:10.1080/08957340701301462

Jennings, J., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. Sociology Of *Education, 87*(2), 125-141. doi:10.1177/0038040714525787

Kelly, A. S., & Orris, J. B. (2011). ASSESSING ACCOUNTABILITY IN U.S. PUBLIC EDUCATION. *Journal Of Public Budgeting, Accounting & Financial Management, 23*(1), 1-30.

Kieffer, M. J. (2008). Catching Up or Falling Behind? Initial English Proficiency, Concentrated Poverty, and the Reading Growth of Language Minority Learners in the United States. *Journal Of Educational Psychology*, *100*(4), 851-868. doi:10.1037/0022-0663.100.4.851

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record, 116*(1), 1-11.

Konstantopoulos, S., & Chung V. (2011). Teacher Effects on Minority and Disadvantaged Students' Grade 4 Achievement. *The Journal of Educational Research, 104,* 73-86. doi:10.1080/00220670903567349

Kotok, S. (2017). Unfulfilled Potential: High-Achieving Minority Students and the High School Achievement Gap in Math. *High School Journal*, *100*(3), 183-202.

Kress, S., Zechmann, S., & Schmitten, J. M. (2011). When Performance Matters: The Past, Present, and Future of Consequential Accountability in Public Education. *Harvard Journal on Legislation*, *48*(1), 185-234.

Lakes, K. D. & Hoyt, W. T. (2009). Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research. *Journal of Clinical Child & Adolescent Psychology*, *38*(1), 144-165. doi:10.1080/15374410802575461

Lee, J. (2010). Trick or treat: new ecology of education accountability system in the USA. *Journal Of Education Policy*, *25*(1), 73-93. doi:10.1080/02680930903377423

Lleras, C., & McKillip, M. (2017). When children move: Behavior and achievement outcomes during elementary school. *Journal Of Educational Research*, *110*(2), 177-187.

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies*, *40*(6), 699-711. doi:10.1080/00220270802105729

Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relationships between High-Stakes Testing Policies and Student Achievement after Controlling for Demographic Factors in Aggregated Data. *Education Policy Analysis Archives*, *14*(30), 1-31.

Maxwell, L. E. (2016). School building condition, social climate, student attendance and academic achievement: A mediation model. *Journal Of Environmental Psychology*, *46*, 206-216. doi:10.1016/j.jenvp.2016.04.009

Murray, K., & Howe, K. R. (2017). Neglecting Democracy in Education Policy: A-F School Report Card Accountability Systems. *Education Policy Analysis Archives, 25*(109), 1-31.

Mushquash, C. & O'Conner (2006). SPSS and SAS Programs for Generalizability Theory Analyses. *Behavior Research Methods, 38*(3), 542-547.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation & Policy Analysis, 26*(3), 237-257.

Overview of the Model to Measure Student Learning Growth on FCAT - Florida Department of Education (2011). Retrieved from http://www.fldoe.org/teaching/performance- evaluation/student-growth.stml

Parke, C. S., & Kanyongo, G. Y. (2012). Student Attendance, Mobility, and Mathematics Achievement in an Urban School District. *Journal Of Educational Research, 105*(3), 161-175.

Peterson, E. R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning And Instruction,* 123-140. doi:10.1016/j.learninstruc.2016.01.010

Peevely, G., Hedges, L., & Nye, B. A. (2005). The Relationship of Class Size Effects and Teacher Salary. *Journal Of Education Finance*, 31(1), 101-109.

Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2015). The effects of scaffolding in the classroom: support contingency and student independent working time in relation

to student achievement, task effort and appreciation of support. *Instructional Science, 43*(5), 615-641.

Rossell, C. H. (2005). The Flawed Requirements for Limited English Proficient Children of the No Child Left Behind Act. *Journal Of Education, 186*(3), 29-40.

School Grades – Florida Department of Education. (2014). Retrieved from http://schoolgrades.fldoe.org/pdf/1314/Guidesheet2014SchoolGrades.pdf

Schulte, A. a., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. T. (2016). Achievement Gaps for Students With Disabilities: Stable, Widening, or Narrowing on a State-Wide Reading Comprehension Test? *Journal Of Educational Psychology, 108*(7), 925-942.

Schwartz, A. E., Stiefel, L., & Wiswall, M. m. (2016). Are all schools created equal? Learning environments in small and large public high schools in New York City. *Economics Of Education Review*, 52272-290.

Scorecard Brief – Michigan Department of Education. (2014). Retrieved from http://www.michigan.gov/documents/mde/Scorecard_Brief_465181_7.pdf

Slama, R. R. (2012). A Longitudinal Analysis of Academic English Proficiency Outcomes for Adolescent English Language Learners in the United States. *Journal Of Educational Psychology, 104*(2), 265-285.

Spera, C., Wentzel, K. R., & Matto, H. C. (2009). Parental Aspirations for Their Children's Educational Attainment: Relations to Ethnicity, Parental Education, Children's Academic Performance, and Parental Perceptions of School Climate. *Journal Of Youth & Adolescence, 38*(8), 1140-1152.

Stevens, J. s., Schulte, A. a., Elliott, S. S., Nese, J. j., & Tindal, G. g. (2015). Growth and gaps in mathematics achievement of students with and without disabilities on a statewide achievement test. *Journal Of School Psychology*, *53*(1), 45-62.

Taylor, J. A. (2005). Poverty and Student Achievement. *Multicultural Education*, *12*(4), 53-55.

Technical Documentation PI Score – Ohio Department of Education (2017). Retrieved from http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Achievement-Measure/Technical-Documentation-PI-Score.pdf.aspx

Toutkoushian, R. K., & Curtis, T. (2005). Effects of Socioeconomic Factors on Public High School Outcomes and Rankings. *Journal Of Educational Research*, *98*(5), 259-271.

Typology of School Districts – Ohio Department of Education (2015b). Retrieved from http://education.ohio.gov/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Typology-of-Ohio-School-Districts

Wang, M., & Degol, J. L. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, *28*(2), 315-352. doi:10.1007/s10648-015-9319-1

Webb, N. M & Shavelson, R. J. (2005). Generalizability Theory: Overview. *Encyclopedia of Statistics in Behavioral Science*, *2*, 717-719.

Wellman, J. V. (2001). Assessing State Accountability Systems. *Change*, 33(2), 47.

White, G. W., Stepney, C. T., Hatchimonji, D. R., Moceri, D. C., Linsky, A. V., Reyes-Portillo, J. A., & Elias, M. J. (2016). The increasing impact of socioeconomics

and race on standardized academic test scores across elementary, middle, and high school. *American Journal Of Orthopsychiatry*, *86*(1), 10-23. doi:10.1037/ort0000122

Wu, Q., Morgan, P. L., & Farkas, G. (2014). Does Minority Status Increase the Effect of Disability Status on Elementary Schoolchildren's Academic Achievement? *Remedial & Special Education*, *35*(6), 366-377.

Young, J. W., Yeonsuk, C., Guangming, L., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and Fairness of State Standards-Based Assessments for English Language Learners. *Educational Assessment*, *13*(2/3), 170-192.