Examining the Effects of Class Size on Student Achievement in the Mohawk Area School District: An Individual Participant Meta-Analytic Investigation

by

Lorree Houk

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Education

in the

Special Education

Program

SLIPPERY ROCK UNIVERSITY

December 2017

EXAMINING THE EFFECTS OF CLASS SIZE ON STUDENT ACHIEVEMENT

I hereby release this dissertation to the public. I understand that this dissertation will be made available from the Bailey Library Circulation Desk for public access. I also authorize the university or other individuals to make copies of this dissertation as needed for scholarly research.

Signature:

11-20-17

Lorree Houk, Student — Date

Approvals:

11/20/17

Dr. Matthew J. Erickson, Dissertation Chair — Date

11|20|17

Dr. Robert Isherwood, Committee Member — Date

11.20.17

Dr. Karen H. Larwin, Committee Member — Date

EXAMINING THE EFFECTS OF CLASS SIZE ON STUDENT ACHIEVEMENT

Abstract

The issue of class size is a contentious issue in American education. While there is substantial empirical support for the existence of a negative correlation between class size and student achievement, numerous practical difficulties prevent schools from being able to further shrink class sizes. Policy-makers and educators need to achieve a more precise understanding of the degree to which class sizes make contributions to student achievement, as only such an understanding can allow decision-makers to set an effective class size policy. Educators have indicated that with smaller class sizes they are able to produce better student achievement results in reading and mathematics. With the increase in teacher accountability, class size is a topic discussed at the school district level.

This independent meta-analytic investigation occurred within the Mohawk Area School District. The study sample included 93 reading classrooms and 97 mathematics classrooms ranging from grade levels kindergarten through eighth grade. The purpose of this quantitative case study based on statistical meta-analysis was to calculate the effect size of class size on student achievement in the Mohawk Area School District, Grades K-8. This purpose was achieved through calculating the effect size using Cohen's *d*. The i-Ready assessment was used as the standardized measure to calculate the effect of class size on achievement in reading and math.

The analysis of this investigation indicates that there is no meaningful effect of class size on reading and math performance. These results are important considerations for the Mohawk Area School District as it faces budget constraints that impact the ability of the district to make class size decisions in the best interest of the staff and students.

Acknowledgements

The guidance and support of the Slippery Rock University faculty has made the completion of this dissertation possible. It was an honor to be a part of the first cohort of doctoral students at SRU. I appreciate the support and input of my committee members, Dr. Robert Isherwood and Dr. Karen Larwin. Dr. Larwin, your expertise in meta-analysis was instrumental in the completion of this study.

Dr. Matthew Erickson, my dissertation chair, was the guiding force behind my success. Your assistance with the planning and construction of this project was invaluable. Dr. E, thank you for your support and encouragement every step of the way.

To the "A" Team at Mohawk who has been my support system throughout my administrative career, a heartfelt appreciation for your daily inspiration throughout this journey.

To Lauren Dillon and Linda Retort, the i-Ready gurus at Mohawk, thank you for your assistance in analyzing our data.

A special note of appreciation to Barb Maravola, my right hand person, who gets me through each and every work day.

Dedications

This dissertation is dedicated to my loving family. You all mean so much to me and have been there supporting and encouraging me through this whole process.

Special appreciation to:

Grammy who takes care of all things and spoils me so that I can pursue my dreams.

My mom who gave me the gift of the appreciation for lifelong learning and the tenacity to accomplish anything at any age.

My dad who instilled in me the fundamental ethics of hard work and dedication. His daily texts and calls mean the world to me.

My sister, Lee Ann, who is my biggest cheerleader and encourager even though we are miles apart.

My children, Evan, Elissa, and Stephen, for being my editors, proofreaders, and making everyday fun! You are my greatest achievements. I love you dearly.

Most importantly, thank you to my best friend, soulmate, and husband, Alan. You have been my number one supporter for over thirty years—my "rock" and the love of my life. I appreciate you immensely and will love you always and forever!

Table of Contents

Chapter 1 1

Introduction 1

Statement of the Problem 2

Purpose of the Study 2

Research Questions 2

Nature of the Study 3

Theoretical Foundation 3

Structure of the Study 4

Conclusion 4

Chapter 2 5

Introduction 5

Overview of Theories 5

Overview of Empirical Studies 8

Historical Overview 8

STAR 12

Recent Empirical Studies 19

Gaps in the Empirical Literature........24

Conclusion........28

Chapter 3........32

Description and Justification of Statistical Meta-Analysis........32

Research Questions........35

Study Variables........37

Measurement........38

Data Collection........40

Reliability and Validity........40

Conclusion........42

Chapter 4........43

Introduction........43

Research Question 1 Results........44

Research Question 2 Results........50

Regression Analysis........55

Conclusion........59

Chapter 5........61

Introduction....61

Statement of the Problem....61

Review of the Methodology....62

Summary of the Results....63

Discussion of the Results....65

Recommendations for Future Research....69

Recommendations for Practice....70

Summary....71

Appendices....72

References....84

Tables

Table 1 *Differences between Successful and Unsuccessful Educational Systems*.....................6

Table 2 *RQ 1 and Sub-RQs of the Study*...36

Table 3 *RQ 2 and Sub-RQs of the Study*...37

Table 4 *Variables in Study*...38

Table 5 *Reading Achievement by Class Size*...45

Table 6 *Reading Achievement by Grade Level*..47

Table 7 *Percentage of Students in Special Education in the Reading Classroom*...................48

Table 8 *Percentage of Students of Low Socioeconomic Status (SES) in the Reading Classroom*.49

Table 9 *Math Achievement by Class Size*...51

Table 10 *Math Achievement by Grade Level*...52

Table 11 *Percentage of Students in Special Education in the Mathematics Classroom*...........53

Table 12 *Percentage of Students of Low Socioeconomic Status (SES) in the Mathematics Classroom*..54

Table 13 *Math Achievement and Reading Achievement Effect Size*..................................55

Figures

Figure 1 *Scatterplot Representation for Reading-Class Size*..........56

Figure 2 *Scatterplot Representation for Reading-Grade Level*..........57

Figure 3 *Scatterplot Representation for Math-Class Size*..........58

Figure 4 *Scatterplot Representation for Math-Grade Level*..........59

Chapter 1

Introduction

In American K-12 education, class size has been an important issue ever since the 19th century, when increasing urban populations created pressure on schools and school districts to be able to concentrate more students into schools. In contemporary American education policy, class size has been conceptualized primarily in terms of educational achievement (Borg, Mary, & Harriet, 2012). On the basis of some evidence (Nye, Hedges, & Konstantopoulos, 2000) that class size is negatively correlated with student achievement, some scholars, policy-makers, educational leaders, and teachers have called for schools to reduce their class sizes. However, economic and administrative pressures on American schools complicate efforts to reduce class sizes. In an environment of declining tax revenues, particularly after the Great Recession of 2008, many school districts find it economically difficult to hire new teachers or to expand the physical infrastructure of their constituent schools.

There appears to be a consensus in the empirical literature that smaller classes are better, as long as factors such as teacher and curriculum quality are controlled for. However, there are still unanswered questions about the relationship between class size and student achievement. One such question has to do with the effect of class size on student achievement. In statistics, there is a distinction (Altman, 1991; Jackson, 2015; Kremelberg, 2010; Moore & McCabe, 2009; Natrella, 2013; Vogt & Johnson, 2011) between statistical significance (also known as a p value) and real-world significance, which can be measured through effect sizes such as Cohen's d (Cohen, 2013). The focus of this quantitative case study based on statistical meta-analysis is the calculation of effect sizes of class size on student achievement in a single American school, the Mohawk Area School District, a study that includes grade levels from K to 8.

Statement of the Problem

The problem is that the relative lack of effect size calculations in the empirical literature does not provide policy-makers, educational leaders, and other relevant stakeholders with the kind of information needed to support decisions to reduce class sizes. In scholarly terms, the problem is rooted in the difference between statistical significance and effect sizes. It is possible for there to be a statistically significant relationship of small class sizes on student achievement, but for the actual magnitude of the effect of class sizes on student class sizes to be too small to warrant changes to class sizes. In practical terms, the problem is that decision-makers lack the kind of data support they need to be able to settle on appropriate class sizes or to set targets for class sizes.

Purpose of the Study

The purpose of this quantitative case study based on statistical meta-analysis is to calculate the effect size of class size on student achievement in the Mohawk Area School District, Grades K-8. This purpose will be achieved through calculating the effect size using Cohen's *d*. A detailed description and justification of this approach to calculating effect sizes will be provided in Chapter 3.

Research Questions

The following research questions guided the study:

RQ1: What is the effect size of classroom size on reading achievement at the school level?

RQ2: What is the effect size of classroom size on math achievement at the school level?

These research questions are not accompanied by hypotheses, as they are descriptive rather than inferential in nature.

Nature of the Study

The study is a quantitative case study based on the use of statistical meta-analysis. The study is quantitative because its main feature is the calculation of a mathematical parameter, an effect size, on the basis of variations in mathematically coded independent, control, and dependent variables. The study is a case study because it is limited to one school district. Finally, the study is meta-analytical in that it includes a pooling of effect sizes from multiple classrooms into single effect size figures. All of these orientations of the study have been further described and justified in Chapter 3.

Theoretical Foundation

The theoretical foundation of the study is the four-factor theory of classroom achievement (Abrantes, Seabra, & Lages, 2007). According to Abrantes et al., a successful classroom is one in which there is (a) a higher quantity and quality of student-teacher interaction, (b) a high degree of responsiveness demonstrated by teachers towards students, (c) a high level of classroom organization, and (d) teacher likeability. There are two main reasons why Abrantes et al.'s theory was chosen as the theoretical foundation for the current study. First, Abrantes et al.'s findings were based on a statistical analysis of the difference between high- and low-achieving classrooms in an environment in which teacher quality and student demographics were already controlled for. Therefore, Abrantes et al.'s four factors of success appear to be genuine differentiating factors between successful and unsuccessful classrooms. Second, Abrantes et al.'s theory is multifactorial and aligns well with several existing theories relevant to classrooms, including theories of warm demanding, classroom management, and self-efficacy, among others. Thus, Abrantes et al.'s four-factor theory represents an empirically supported blending of pre-existing theories whose explanatory powers have already been established in the literature.

Structure of the Study

The study has been structured as follows. The first chapter introduced the problem, which was the lack of insight into the effect size of class sizes on student achievement, resulting in a lack of appropriate decision support for policy-makers and educational leaders tasked with setting class size policies. The second chapter, the review of literature, contains an overview of theoretical themes and empirical findings relevant to the study. The third chapter provides a discussion and defense of the methodology and research design of the study. The fourth chapter contains the findings of the study, with particular emphasis placed on the calculated effect size. The fifth and concluding chapter discusses the findings in light of existing theories and empirical literature, acknowledges the study's limitations, and presents recommendations for both practice and future research on the topic of class sizes and academic achievement.

Conclusion

The issue of class size remains a contentious issue in American education. While there is substantial empirical support for the existence of a negative correlation between class size and student achievement, numerous practical difficulties prevent schools from being able to further shrink class sizes. Policy-makers and educators need to achieve a more precise understanding of the degree to which class sizes make contributions to student achievement, as only such an understanding can allow decision-makers to set an effective class size policy.

Chapter 2

Literature Review

Introduction

The problem identified in Chapter 1 was the absence of insight into the effect sizes of class size on academic achievement. The purpose of the literature review is to discuss both theoretical themes and empirical findings that are relevant to this problem. The theoretical discussion is based primarily on Abrantes et al.'s (2007) four-factor theory of classroom success. The review of empirical studies focuses on quantitative studies of the effect of class size on academic performance. A brief conclusion provides a synthesis of the research findings and explanatory themes.

Overview of Theories

According to Henderikus, a theory "is normally aimed at providing explanatory leverage on a problem, describing innovative features of a phenomenon or providing predictive utility" (Henderikus, 2010, p. 1498). Thus, in order to be effective, the theoretical background of the current study should possess explanatory power, predictive utility, and descriptive robustness. The chosen theoretical framework for this study was the four-factor theory of successful classrooms (Abrantes et al., 2007). Because the four-factor theory is highly important in the context of this study, it has been described in detail in Table 1 below, which is based on Abrantes et al.'s description (Abrantes et al., 2007, pp. 961-962) of successful and unsuccessful classrooms. As discussed subsequently in this chapter, Abrantes et al.'s four-factor model aligns well with previous theories, particularly those of warm demanding, self-efficacy, classroom management, and engagement.

Table 1

Differences between Successful and Unsuccessful Educational Systems

Variable	Successful System	Unsuccessful System
Student-teacher interaction	Higher quantity of interactions; qualitatively, interactions will motivate, energize, and inform the student.	Lower quantity of interactions; qualitatively, interactions will demotivate, demoralize, and fail to inform the student.
Responsiveness	Both the teacher and the system will respond well to the student's individual needs by demonstrating curricular and pedagogical flexibility, creativity, and individualization.	Neither the teacher nor the educational system will respond well to the student's individual needs by demonstrating curricular and pedagogical flexibility, creativity, and individualization.
Organization	Both the teacher and the educational system will demonstrate appropriate selection, presentation, and pacing of curricular materials in a manner likely to achieve the goals of education.	Neither the teacher nor the educational system will demonstrate appropriate selection, presentation, and pacing of curricular materials in a manner likely to achieve the goals of education.
Teacher Likeability / Concern	The teacher will be likeable, and will demonstrate warm, demanding behaviors to the student.	The teacher will not be likeable, and will not demonstrate warm, demanding behaviors to the student.

Abrantes et al.'s (2007) four-factor model is a plausible rationale for why there might be a relationship between smaller classrooms and improved academic performance. Abrantes et al.'s variable of student-teacher interaction is clearly dependent upon classroom size, as, holding other factors equal; classrooms that are smaller in size are more likely to allow teachers to increase their interactions with each student. The Abrantes et al. variable of responsiveness is also conceptually related to classroom sizes, as, once again holding other factors equal, it seems that teachers will be able to improve their responsiveness to individual students (in terms of pedagogical and curricular customization, for example), when there are fewer students amongst whom the teacher's attention has to be divided. The Abrantes et al. variable of organization touches upon classroom management, suggesting that classroom organization would be easier if the teacher is not overwhelmed by the demands of a large classroom. Finally, the Abrantes et al. variable of teacher likeability and concern also appears to be conceptually related to classroom sizes, insofar as teachers may be more readily able to express concern to smaller classrooms. Because teachers might be more likely to experience burnout, frustration, and other negative psychological affective states in larger classrooms, classroom size may have the ability to reduce teacher likeability and concern.

Thus, each of the Abrantes et al. (2007) variables and the model in general, appears to (a) predict the existence of a positive relationship between smaller class sizes and improved academic performance and (b) provide a rationale for the existence of such a relationship. Thus, Abrantes et al.'s model appears to possess the main characteristics of a success theory as defined by Henderikus (2010). Abrantes et al.'s model provides predictive utility, as this model suggests that there will be a statistically significant impact of class size on academic performance. Abrantes et al.'s model also provides what Henderikus referred to as "explanatory leverage" (Henderikus, 2010, p. 1498), insofar as the variables in Abrantes et al.'s model provide plausible explanations for why a positive relationship between lower class sizes and academic performance might exist. Finally, Abrantes et al.'s variables also describe the features of teacher-student interaction in terms of factors that are related to class size, providing an overall context in the effect of classroom size on academic performance which can be investigated.

Overview of Empirical Studies

The purpose of the review of empirical studies is to describe, critically evaluate, and synthesize past empirical results related to the topic of classroom size and academic achievement. The empirical results have been divided into several categories. The first category includes an overview of the early and seminal results on this topic. The second category is devoted to the analysis of the Student/Teacher Achievement Ratio (STAR) pseudo-experiment related to classroom size and academic performance. The third category is devoted to an overview of more recent (2011-onwards) empirical studies on the relationship between classroom size and academic success.

The inclusion of some older studies and pseudo-experimental results requires some further justification. In the United States in particular, there have been numerous attempts to reduce classroom size and measure the impact, if any, on student performance, and several of these attempts date back to the 1980s. For example, Student-Teacher Achievement Ratio (STAR) took place from 1985 to the early 1990s. Discussing the results of STAR and other important pseudo-experiments related to class size and academic performance therefore requires a discussion of some studies from the 1980s and 1990s that reported on pseudo-experimental results.

Historical Overview

The issue of class size is quite old and has been examined in some fascinating historical contexts. The 12^{th}-century Jewish scholar, Moses Maimonides, issued a recommendation that class sizes be capped at 40 students (Cho, Glewwe, & Whitler, 2012). Although this injunction is nearly a millennium old, it is still the backbone of class size policy in the state of Israel, in which no public school classroom contains more than 40 students (Cho et al., 2012). In the United States, class size appears to have become an important issue towards the end of the 19th century, when the combining forces of urbanization and immigration resulted in the concentration of large numbers of students, often students who lacked English knowledge and other prerequisites of education, in American classrooms (Go & Lindert, 2010). However, changes in American immigration policy in the 1920s (Gans, Replogle, & Tichenor, 2008; Hing, 1994; Ngai, 1999) lowered the number of immigrants, and the combined forces of assimilationist social pressures and increasingly standardized school approaches meant that, for several years, the issue of classroom sizes did not arise again.

The next major historical development with implications for classroom sizes was the United States Supreme Court's pivotal 1954 decision in *Brown v. School Board of Topeka*, which resulted in the racial integration of American schools. Until that pivotal decision, American schools that served black children had higher classroom sizes, as the budgets of such schools were lower than those of schools that served white children (Hartney & Flavin, 2014). The pre-*Brown* disparity in classroom sizes between schools that served white children and schools that served black children is a unique illustration of the power of classroom sizes to positively impact academic performance. While there do not appear to have been pre-1954 empirical studies on the relationship between classroom sizes and academic achievement, the fact that white schools had smaller classroom sizes indicates that educational leaders were already aware of the benefits of smaller classroom sizes, benefits that were disproportionately routed to white schoolchildren.

After *Brown*, issues of classroom size rose to the forefront because of the need to integrate what had been the large classrooms of schools that served black children with the smaller classrooms of schools that served white children (Anderson, 2012; Borg et al., 2012; Burchinal et al., 2011; Cummins, 2009; C. Davis, 2012; Durham, 2012; Hall Mark, 2013; Rios, 2012; Rojas-LeBouef & Slate, 2012). In addition, new changes to immigration laws in the 1960s increased immigration to the United States, as previous country quotas were eliminated or altered, and the process of residence acquisition for the foreign relatives of American green card holders or naturalized citizens also became simpler (Gans et al., 2008; Hing, 1994; Ngai, 1999). In addition, the phenomenon of urbanization continued to increase pace. All of the above factors led to larger classroom sizes in the America of the 1960s and afterwards.

The first quantitative attempts to measure the impact of classroom sizes on academic achievement appear to date from the 1970s, after which such studies began to be more common in the literature. A study from 1980 (Shapson, Wright, Eason, & Fitzgerald, 1980) was one of the first well-controlled pseudo-experiments on the relationship between classroom size and academic performance. This study was carried out on 62 third- and fourth-grade classes distributed among three school districts in Toronto and included the random assignment of students to four class sizes: 16 students, 23 students, 30 students, and 37 students. Shapson et al.'s research design was longitudinal, allowing the researchers to track the outcomes at the end of a two-year period of exposure to the different class sizes. Shapson et al. found a statistically significant relationship between exposure to the smallest (16-student) class and mathematics performance at the end of two years, but the effect size of this relationship was not quantified, and the research design also failed to include several covariates, such as teacher experience, that could also have explained the relationship between classroom size and academic performance.

By 1982, there appeared to be enough empirical studies on the topic of the relationship between classroom size and academic achievement for Glass (Glass, 1982) to publish a book-length meta-analysis and systematic review. Glass reached the following five conclusions, each of which represented an important contribution to the then-emerging scholarly consensus on the topic of classroom size and academic achievement:

> (1) class size is strongly related to pupil achievement; (2) smaller classes are more conducive to improved pupil performance than larger classes; (3) smaller classes provide more opportunities to adapt learning programs to individual needs; (4) pupils in smaller classes have more interest in learning; (5) teacher morale in smaller classes is better; (6) smaller classes improve pupil achievement because they allow teachers to spend more time with individual students; (7) smaller classes lead to increased student attention. (Glass, 1982, p. 1).

The results of Glass's (1982) meta-analysis were replicated or triangulated in a substantial number of future studies, including studies discussed in the remainder of the literature review. However, not all of the early studies on the relationship between classroom size and academic achievement supported the conclusion that smaller class sizes had a positive effect on academic performance. One of the most influential empirical articles on the relationship was that of Hoxby (Hoxby, 2000). Hoxby performed a correlational analysis of the relationship between classroom size and academic performance and found no statistically significant effect. However, Hoxby assumed that the distribution of class sizes was random, whereas, as Mueller (2013) pointed out, students with higher impediments to academic performance (such as diagnosed behavioral difficulties) are often purposely assigned to smaller classes. Therefore, variation in class sizes is unlikely to be as random as Hoxby assumed, which, in turn, means that the absence of a relationship between classroom size and academic performance identified in Hoxby's work could be an unreliable finding.

Historically, the quantitative literature on classroom size and academic performance can be separated into two periods, the pre-STAR period and the post-STAR period. Because STAR was the first—and, to date, the only—large-scale experimental pseudo-experiment on the relationship between classroom sizes and academic achievement conducted at either the state or the federal level in the United States. The scale of STAR, and its unique nature, means that much of the discussion of classroom sizes and academic achievement in the United States has referred to STAR in some manner.

STAR

Before reporting on STAR results, it should be noted that the STAR project released its results and datasets publicly, allowing scholars to report on the findings on their own. As a result, the reporting of the STAR results that follows in this section is not indebted to any specific set of scholars, but rather to the raw data and analysis presented in several studies (Hanushek, 1999; Mosteller, 1995; Mueller, 2013) that have included discussions of STAR data. Because the STAR data are extensive, they have been utilized in many ways. For example, Mueller (2013) used the STAR data for the purpose of quantile regression, which is an advanced statistical technique. The purposes of this section of the literature review are to (a) report on the more basic aspects of the STAR study and (b) more closely examine some of the seminal interpretations of the STAR results.

Tennessee's STAR project was probably the earliest, and among the largest, pseudo-experiments related to the measurement of the effect of classroom sizes on academic performance. This experiment began in 1985 and ended in 1990. During this time, 11,600 younger (grades K-3) students in Tennessee were randomly assigned to classrooms of varying size. The pseudo-experiment was carried out in 17 school districts in Tennessee that had been identified as having a particularly high dropout rate.

Before discussing the results of the STAR study further, the distinction between pseudo-experiments and experiments requires clarification. One of the characteristics of a true experiment is the ability to randomly assign members of a sample to a control group and a treatment group or groups (Zhang, Zhang, & Seiler, 2013). The purpose of doing so is to obtain a more valid measurement of the impact of an intervention, such as that of a student being moved to a classroom of smaller size or on an outcome of interest, such as academic performance. STAR and similar interventions have been described as experiments by some scholars (Hanushek, 1999; Mosteller, 1995), because these interventions have featured random assignments of participants. However, there is a consensus among experts on experimental methodology that a true experiment requires a controlled (typically, a laboratory-based) environment (Creswell, 2015; Jackson, 2015; Leary, 2011; McBurney & White, 2011; Trochim, Donnelly, & Arora, 2015; Zikmund, 2003). In projects such as STAR, there was no way to rule out confounding effects that took place outside schools; for example, if parents who found out that their children were assigned to large classrooms decided to respond by increasing the amount of tutoring or academic assistance that they were providing their children, then any observed impact of classroom size on academic performance might have been confounded. STAR and similar studies are better described as pseudo-experiments rather than experiments, because, despite featuring random assignment, they took place in real-world conditions in which students, parents, teachers, and others all had the ability to confound the results of intervention in the manner described above.

In the STAR intervention, there were three conditions. The first condition was a classroom that consisted of between 13 to 17 students. This classroom was designated as being small. The second condition was a classroom that consisted of between 22 to 25 pupils. This classroom was designed as being of regular size. The third condition was a classroom that consisted of between 22 to 25 pupils, but that also included a teacher's aide. Conceptually speaking, the classrooms that consisted of 22 to 25 students and that had no teacher's aide can also be described as large, and some scholars have used this designation. One reason for this ambiguity in terminology is that, in Tennessee, it was common for K-3 students to be in classes that contained up to 24 others; thus, the normal size was that of a large classroom.

The STAR intervention was longitudinal and drew upon a matched-pairs approach. In the first phase of the study, which lasted from 1985 to 1989, K-3 students in 17 high-dropout-risk school districts in Tennessee were randomly assigned to one of the three classroom conditions specified above (small, large, and large with a teacher's aide). Afterwards, the children in the study entered large classrooms. Thus, the STAR intervention was designed to measure the impact of exposure to smaller classrooms in grades K-3 on the academic performance of children in the fourth, fifth, sixth, and subsequent grades.

The main finding of the STAR study was that having been in smaller classes during grades K-3 was associated with higher academic performance in subsequent years, starting in the fourth grade. Students who had been in smaller classes during the STAR intervention had significantly higher reading and mathematics scores than either students in large classes or students in large classes with teachers' aides. In addition to measuring statistical significance, scholars who examined the results of the STAR study also calculated effect sizes for the impact of smaller classes on academic performance. It was found that the effect size of being in a smaller class, as compared to being in a larger class or a larger class with a teacher's aide, was 0.23 for reading and 0.27 for math (Hanushek, 1999; Mosteller, 1995). Although interpreting the magnitude of effect sizes is highly reliant on what is being measured and on other context factors (Cohen, 2013), the effect sizes obtained in the STAR can be described as being on the lower side of moderate. Cohen, an authority on effect sizes, has stated that an effect size of 0.2 is small, an effect size if 0.5 is medium, and an effect size of 0.8 is large. Because the effect sizes obtained in the STAR study were greater than 0.2, but well short of 0.5, they can be described as being on the lower end of medium or on the higher end of small.

However, the magnitude of the STAR effect sizes has to be considered in light of the study design. Because of the combination of random assignment and the inclusion of thousands of children in the STAR study, it was likely that the effect sizes are not spurious and represent a genuine effect of the intervention (Hanushek, 1999; Mosteller, 1995). In this context, the achievement of even a modest improvement in performance is notable. There is a consensus among psychologists that general intelligence is a fixed quantity that is inherited and that is subject to minimal modification over the lifespan. Therefore, the achievement of even low-moderate improvements in reading and math as a result of exposure to smaller class sizes is meaningful.

In the United States, the STAR pseudo-experiment remains the only experiment of its scale conducted to date. Because this experiment was conducted only in at-risk districts in a specific state, and because its results are several years old, there are plausible reasons to critique the generalizability of STAR findings. However, recent empirical studies, although conducted on much smaller samples than those in the STAR study and often failing to include random assignment, have reached much the same conclusions as STAR, providing further confirmation for the claim that lower class sizes are associated with higher levels of academic performance.

As noted earlier in this section, the STAR results were made public, and numerous scholars have drawn upon this data to generate findings related to the STAR pseudo-experiment. The purpose of the following paragraphs is to discuss how the STAR results were reported in some of the early, seminal work associated with the STAR project. This early work is of interest because of its reporting of effect sizes and basic insights, but it is also limited by the absence of more complex and explanatorily powerful statistical techniques.

Hanushek's (1999) report on the STAR results was of particular interest for the author's joint reporting of the STAR results and the results of several other interventions designed to improve academic performance by reducing classroom sizes. Initially, Hanushek carried out a statistical meta-analysis designed to summarize all available studies on the topic of class sizes and academic achievement. As part of this meta-analysis, Hanushek tabulated previous results on classroom size and academic achievement in terms of whether they were statistically significant or insignificant, whether they found positive or negative effects of lower classroom size on student performance and by school level (including the levels of all schools, elementary schools only, and secondary schools only).

Hanushek (1999) tabulated the results from 277 studies, including STAR, in this initial tabulation of effects. However, the studies were not limited to the United States, which casts doubt on their ability to reliably capture the classroom size-academic performance as it exists in the U.S. In addition, Hanushek did not make an attempt to weight studies by the quality of the study; all studies were given equal weight in Hanushek's meta-analysis. In the initial tabulation, Hanushek found that the existing studies appeared to be equally split concerning the effects of classroom size on academic performance, with a roughly equal number of (a) studies having found that smaller classroom sizes were associated with improved academic performance and (b) studies having found that smaller classroom sizes were associated with reduced academic performance. After presenting this meta-analysis, Hanushek proceeded to a discussion of STAR findings.

Hanushek's (1999) attempt at meta-analysis illustrates some of the gaps in the early meta-analytical literature on classroom size and academic performance. One of the points of consensus in the current literature (Bonta, Blais, & Wilson, 2014; Burton et al., 2015; Clark, Tanner-Smith, & Killingsworth, 2016; Cole & Dendukuri, 2014; Cuijpers, Weitz, Karyotaki, Garber, & Andersson, 2015; Franz, Boucher, Rutten-Ramos, & VanWormer, 2015; Luebke & Brunkwall, 2015; Madigan, Atkinson, Laurin, & Benoit, 2013; Nunnery, Chappell, & Arnold, 2013; Schmid et al., 2014; Tobias et al., 2015; Verhulst, Neale, & Kendler, 2015; Voyer & Voyer, 2014) on statistical meta-analysis is that the main purpose of meta-analysis is to calculate an effect size. The theory behind effect size pooling is that, while single studies might be flawed or otherwise unreliable, the collective effect size emerging from the synthesized results of several studies is more likely to reflect the true impact of an intervention on a variable of interest (Bonta et al., 2014; Burton et al., 2015; Clark et al., 2016; Cole & Dendukuri, 2014; Cuijpers et al., 2015; Franz et al., 2015; Luebke & Brunkwall, 2015; Madigan et al., 2013; Nunnery et al., 2013; Schmid et al., 2014; Tobias et al., 2015; Verhulst et al., 2015; Voyer & Voyer, 2014). While Hanushek did tabulate the results of 277 previous studies in terms of whether these studies found a positive or negative effect of smaller classroom sizes on academic performance, the failure to calculate a common effect size across these studies, whether through simple averaging or a weighted method, was an important limitation of the study (Bonta et al., 2014; Burton et al., 2015; Clark et al., 2016; Cole & Dendukuri, 2014; Cuijpers et al., 2015; Franz et al., 2015; Luebke & Brunkwall, 2015; Madigan et al., 2013; Nunnery et al., 2013; Schmid et al., 2014; Tobias et al., 2015; Verhulst et al., 2015; Voyer & Voyer, 2014).

Hanushek's (1999) analysis of the STAR data had the same limitation. Hanushek, unlike Mosteller (1995) did not calculate effect sizes for any improvement data related to STAR. Based on the meta-analytical literature (Bonta et al., 2014; Burton et al., 2015; Clark et al., 2016; Cole & Dendukuri, 2014; Cuijpers et al., 2015; Franz et al., 2015; Luebke & Brunkwall, 2015; Madigan et al., 2013; Nunnery et al., 2013; Schmid et al., 2014; Tobias et al., 2015; Verhulst et al., 2015; Voyer & Voyer, 2014), effect sizes are calculated in terms of Cohen's *d*.

Mosteller's (1995) study also reported on the STAR findings. However, unlike Hanushek (1999), Mosteller reported on the effect sizes of STAR. As a statistician, Mosteller appeared to be more familiar than Hanushek with the calculation and reporting of effect sizes. Moreover, as an expert on statistics, Mosteller provided important context for the small-to-moderate effect sizes calculated in STAR. Mosteller wrote that:

> Although effect sizes of the magnitude of 0.1, 0.2, or 0.3 may not seem to be impressive gains for a single individual, for a population they can be quite substantial. For example, a 0.2 effect size corresponds in the United States to the difference between the average heights of 15-year-old versus 16-year-old girls. For large numbers of girls of each age, this average difference may seem small, but most people notice it. An effect size of 0.3 corresponds to about 30 points on a SAT verbal or mathematics standardized test. (Mosteller, 1995, p. 120).

Mosteller's (1995) paper was published some years after the conclusion of the STAR study. However, Mosteller appears to have been the first scholar to have calculated the effect sizes of performance improvement associated with the small class exposure group in STAR. Moreover, as noted above, Mosteller provided an interpretative context for the calculated effect size that establishes the real-world importance of the STAR intervention. Therefore, Mosteller's paper ought to be considered among the most important empirical papers to have reported the STAR findings.

Recent Empirical Studies

Numerous recent empirical studies have found a statistically significant effect of smaller class sizes on improved academic performance, particularly in terms of reading and mathematics scores. A study carried out by Cho et al. (2012) found that there was a statistically significant impact of classroom size on reading and mathematics achievement in several elementary schools in Minnesota. The premise of Cho et al.'s paper was that the impact of classroom size on classroom achievement can be measured in a correlational manner—that is, in a manner reflecting the existing, natural, and random distribution of classroom sizes. Using this correlational approach, Cho et al. found the existence of a statistically significant, but, in terms of effect size, a modest impact of classroom sizes on academic performance. Specifically, Cho et al. found that a decrease of 10 students increased reading and mathematics standardized test scores among Minnesotan students in grades 3 and 5 by between 0.04 and 0.05 standard deviations. One limitation of Cho et al.'s study in comparison to pseudo-experimental results from STAR and other interventions is that Cho et al. did not model the longitudinal impact of exposure to smaller class sizes. In STAR, the pseudo-experimental exposure to smaller class sizes lasted several years; students were not allowed to enter larger classrooms until the fourth grade. However, in Cho et al.'s correlational approach, the data were cross-sectional, not longitudinal. Thus, even though Cho et al. found that there was a statistically significant and positive effect of smaller classroom sizes on higher levels of achievement, it is possible that the effect would have been even greater had Cho et al. found a way to control for multi-year exposure to smaller class sizes.

Mueller carried out a study on the mediating role played by teacher experience in the STAR results (Mueller, 2013). Mueller's study was notable not only for its inclusion of the mediating role of teacher experience but also because of the use of quantile regression. As Mueller explained, ordinary least squares (OLS) regression attempts to find a line of best fit across all data values. Quantile regression, on the other hand, applies regression at different quantiles of a dependent variable such as achievement. One common approach to quantiles is declines—that is, breaking up a value into 10 equal percentiles. In Mueller's approach, the impact of small class sizes, inexperienced (rookie) teachers, and rookie teachers in small classrooms was considered through the application of both OLS and quantile regression models.

Mueller (2013) utilized 9 quantiles, namely the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentiles of achievement in both reading and mathematics. Mueller's use of an OLS model indicated that having a small classroom size improved reading by 13.6% over having a large classroom size. However, the use of quantile regression indicated that, on the whole, the positive effect of smaller classroom sizes was greater for higher levels of achievement. For example, at the 10th percentile of reading achievement, being in a small classroom improved reading scores by 8.4% over being in a large classroom. However, at the 80th percentile of reading achievement, this difference was 17.5%. Thus, Mueller succeeded in demonstrating that the impact of smaller classroom sizes on reading achievement is greater at the higher levels of achievement than at the lower levels of achievement.

Mueller's (2013) regressions on mathematics achievement found results that were somewhat different from the quantile regressions on reading achievement. Mueller's use of an OLS model indicated that having a small classroom size improved mathematics by 16.2% over having a large classroom size. The use of quantile regression indicated that, on the whole, the positive effect of smaller classroom sizes was similar at all levels of mathematics performance, with the exception of the 10th percentile of achievement, where the positive effect of small classrooms was an 11.5% greater score.

Mueller (2013) also analyzed the impact of rookie teachers and rookie teachers in small classrooms on reading and mathematics achievement. Mueller found a negative effect of rookie teachers in small classes on every quantile in the model and also in the OLS regression. These findings can be interpreted as suggesting that the positive impact of small classrooms on mathematics and reading achievement is dependent on the presence of experienced teachers, because, once teacher experience is factored into the STAR data, the impact of smaller classes on achievement becomes negative rather than positive. Thus, experienced teachers appear to be a necessary component of the ability of smaller class sizes to yield improved academic achievement. This finding is aligned with other empirical findings (Bonner, 2014; Moyer-Packenham, Salkind, Bolyard, & Suh, 2013; Salkovsky & Romi, 2015) that experienced teachers are more skilled at providing customized attention to students.

Given the formation of a consensus (Cho et al., 2012; Glass, 1982; Go & Lindert, 2010; Hanushek, 1999; Hoxby, 2000; Matta, Guzman, Stockly, & Widner, 2015; Mosteller, 1995; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980) that smaller classroom sizes are associated with superior academic performance, scholarly attention has begun to shift to questions of why this relationship might exist. One possible explanation is the variable of parental engagement (Rodriguez & Elbaum, 2014). Rodriguez and Elbaum found that, among the variables of classroom size, school socioeconomic status, and grade level, classroom size was the strongest predictor of parental engagement. Rodriguez and Elbaium hypothesized that, because parental engagement is positively correlated with student achievement (Fan & Williams, 2010; Green et al., 2012; Griner & Stewart, 2013; Hampden-Thompson, Guzman, & Lippman, 2013; Houchen, 2013; Plunkett, Behnke, Sands, & Choi, 2009; Sun, 2014), one mechanism through which smaller classroom sizes impact academic performance is through the promotion of parental engagement.

There is also a substantial body of literature (Ahmed, Minnaert, Kuyper, & van der Werf, 2012; Blondal & Adalbjarnardottir, 2012; Bouta, Retalis, & Paraskeva, 2012; Ghazvini, 2011; Green et al., 2012; Jeynes, 2012; Lee, Hayes, Seitz, DiStefano, & O'Connor, 2016; Mo, Singh, & Chang, 2013; Price & Tovar, 2014; Senler & Sungur, 2009; Usher & Pajares, 2009) that identifies student engagement as one of the main mechanisms through which lower classroom sizes might be able to improve academic performance. Student engagement has been described in terms of not only the student's innate self-concept and orientations towards academic work but also in terms of the teacher's ability to motivate the student (Ahmed et al., 2012; Blondal & Adalbjarnardottir, 2012; Bouta et al., 2012; Ghazvini, 2011; Green et al., 2012; Jeynes, 2012; Lee et al., 2016; Mo et al., 2013; Price & Tovar, 2014; Senler & Sungur, 2009; Usher & Pajares, 2009). Insofar as smaller classrooms afford teachers added opportunity to reach out to engage students in effective ways, student engagement ought to be taken into consideration as a potential explanatory factor.

Another possible rationale for why smaller class sizes might improve student engagement is the theory of warm demanding (Bondy, Ross, Hambacher, & Acosta, 2013; Bonner, 2014;

Ford & Sassi, 2014; Houchen, 2013; Ross, Bondy, Gallingane, & Hambacher, 2008; Xu, Coats, & Davidson, 2012). Bondy et al. defined warm demanding (Bondy et al., 2013, p. 27) as the following set of teacher orientations toward students:

- Conveying personal interest in students
- Understanding how students perceive care (for example, in varying cultural contexts) and ensuring that demonstrated care is communicable and meaningful to the student
- Working hard to remain committed to the welfare of students
- Ensure that ordinary actions (such as pedagogy, gestures, classroom management, etc.) are utilized at least as frequently as verbal affirmation in order to convey care
- Being insistent in bringing students back from digressions to their assigned work
- Putting academic work before ‘fun’ activities and breaks
- Reminding students that they are capable of high achievement
- Holding students accountable for high achievement

Although warm demanding does not appear to have been studied in the context of statistical analyses of the relationship between classroom size and academic performance, the existing empirical work on warm demanding (Bondy et al., 2013; Bonner, 2014; Ford & Sassi, 2014; Houchen, 2013; Ross et al., 2008; Xu et al., 2012) suggests that the process of communicating warmth, and of expressing accountability, is a time- and resource-intensive process, which, in turn, suggests that warm demanding might be substantially easier to achieve in smaller classrooms.

Thus, studies on both teacher and student qualities suggest that the positive academic effects of smaller classrooms appear in a particular context. It is not merely the reduction in the number of students in a class that is efficacious. Rather, the reduction of students appears to give both teacher and students the opportunities to interact and behave in a manner that is more predictive of academic success.

Gaps in the Empirical Literature

There are several gaps in the empirical literature. The existence of such gaps is rendered inevitable by the vast number of explanatory variables involved in any analysis of the relationship between classroom size and academic performance. While the relationship between classroom size and academic performance can be, and has been, quantified through the use of many statistical approaches (Cho et al., 2012; Glass, 1982; Go & Lindert, 2010; Hanushek, 1999; Hoxby, 2000; Matta et al., 2015; Mosteller, 1995; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980), the reasons for the existence of this relationship have not been comprehensively modeled in existing empirical studies. Ideally, empirical analyses would include not only a statistical measure of the effect of classroom size on student performance but also measurements of explanatory teacher variables (such as morale, engagement, and self-efficacy), explanatory student variables (such as engagement, self-concept, and self-efficacy), and environmental variables. Several of the studies reviewed in this chapter have included measurements of the contributions of some of these explanatory variables, but none of the reviewed studies included a mixture of teacher-, student-, and environment-related variables that cumulatively help to explain why the relationship between smaller class sizes and improved academic performance exists. Thus, one of the main gaps in the empirical literature on the topic of the relationship between classroom size and academic performance is the failure to include control variables and covariates capable of existing why smaller classroom sizes are so reliably associated with improved academic performance.

The purpose of the current study is not to close this gap in the literature by including a large number of covariates and engaging in a longitudinal analysis of performance, but to calculate the relationship between classroom sizes and academic performance in a cross-sectional, correlational manner. Therefore, it is important to demonstrate gaps in the existing approaches to measuring the relationship between classroom size and academic achievement that can be closed by the current study. There are two such gaps that are of particular interest in this context: (a) gaps related to the inclusion of certain explanatory variables and (b) gaps related to grade-specific effects.

Several studies (Cho et al., 2012; Glass, 1982; Go & Lindert, 2010; Hanushek, 1999; Hoxby, 2000; Matta et al., 2015; Mosteller, 1995; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980) on the relationship between classroom sizes and academic performance have either failed to control for key explanatory variables or have included too few explanatory variables in statistical analysis. Five key variables of this kind include: (a) the teacher's years of experience, (b) the teacher's gender, (c) the percentage of students in special education, (d) the percentage of students of low socioeconomic status, and (e) the amount of minutes spent by paraprofessionals in the classroom. Some of these variables have indeed been included in previous empirical studies. However, all of these variables do not appear to have been included in any empirical study reviewed in this chapter.

Cumulatively, the five explanatory variables listed above offer potential explanations of why there could be a positive impact of smaller classroom sizes on academic performance. For example, it could be the case that more experienced teachers are purposely assigned to smaller classes, that teachers of a particular gender are more adept at manning classrooms of a certain size, that special education students or students of low socioeconomic status benefit disproportionately from smaller classroom sizes, or that classroom size becomes less of an explanatory factor when controlling for the presence of a paraprofessional or other co-teaching presence.

In fact, there is support for several of these possibilities in the empirical literature, albeit in studies not on the topic of classroom size and academic performance. Scholars have demonstrated that (a) more experienced teachers are better at managing larger classrooms, which means that teacher experience could mediate the relationship between class size and the academic performance of students (Adkins-Coleman, 2010; Bondy et al., 2013; Gordon, 2012; Grantham & Biddle, 2014; Moyer-Packenham et al., 2013; Mueller, 2013; Phillippo, 2012); (b) special education students benefit from smaller classes, in which more customization and specialized attention are possible (Borg et al., 2012; Morrison et al., 2013; Salvia, Ysseldyke, & Bolt, 2012); (c) students of low socioeconomic status also tend to benefit from more customization of the learning environment and from the more focused attention of the teacher (Bonner & Adams, 2012; H. A. Davis, Gabelman, & Wingfield, 2011; Dodson, 2012; Durham, 2012; Hayes, 2012; Theron, 2013; Williams & Sánchez, 2012; Xu et al., 2012), both of which are more likely to occur in smaller classes; and (d) co-teachers make it easier to increase the amount of individualized attention that each student is able to receive (Bonner, 2014; Howes & Ritchie, 2002; Roberts & Dyer, 2004). This empirical evidence is important because it suggests that there are sound reasons for including the covariates of the teacher's years of experience, the percentage of students in special education, the percentage of students of low socioeconomic status, and the amount of minutes spent by paraprofessionals in the classroom in any statistical analysis of the relationship between classroom size and academic achievement. The inclusion of these covariates is therefore likely to close an important gap in the existing empirical literature on classroom size and academic achievement.

Another important gap in the existing empirical literature (Cho et al., 2012; Glass, 1982; Go & Lindert, 2010; Hanushek, 1999; Hoxby, 2000; Matta et al., 2015; Mosteller, 1995; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980) on the relationship between classroom size and academic performance is the failure to provide analyses for multiple grade levels. There are many reasons for the existence of this gap. One reason is that, in studies that report on pseudo-experimental findings, the statistical analysis is necessarily restricted to those grades in which a classroom size reduction intervention has taken place. If such an intervention is limited to a single grade, then research on such an intervention is also limited to a single grade.

There are innate problems of reliability related to testing the effects of classroom size on academic performance on single grades or even on a few grades. In statistics, *reliability* refers to the ability of a test or measurement to return the same results, or acceptably similar results, over several administrations (Altman, 1991; Csörgo & Révész, 2014; Jackson, 2015; Kremelberg, 2010; Moore & McCabe, 2009; Natrella, 2013; Vogt & Johnson, 2011). It is possible that, for a certain class setting examined in a particular year, the measured impact of classroom size on academic performance is the result of accident, or of factors that are not conceptually or operationally related to the variable of class size. On the other hand, measuring the relationship between classroom size and academic performance in the context of several classes (for example, K-8 classes, analyzed separately) increases the reliability of overall conclusions related to the relationship between classroom sizes and academic performance. In such an analysis, it might be found that there is a positive impact of smaller classroom size on academic performance in only 1 of the 9 measured grades. If a scholar chose only 1 out of 9 grades to subject to an effect size analysis, and if that class were the only class in a given school in which there actually was a positive effect of smaller classroom size on academic achievement, then the finding might be unreliable, as it could reflect accidental factors. On the other hand, if all 9 grades in a given school were analyzed, and if the existence of a positive relationship between classroom size and academic performance could be discerned in several of the classes, then the findings would be more reliable, as repeated administrations of effect-size calculations would have found the same conclusion in different grades.

In addition to concerns related to reliability, there are other empirical problems associated with the selection of a single grade or of a limited number of grades on which to conduct a statistical analysis of the relationship between classroom size and academic performance. One such problem is that of discriminant validity. For example, if the positive relationship between classroom effect sizes and academic performance exists only for younger students, then the failure to test the effects of classroom size on academic performance in certain blocks of grades (for example, elementary school, middle school, and high school) would mean an inability to establish the relationship between student age and the classroom size-academic performance effect. Therefore, conducting studies of the relationship between classroom size and academic performance in which multiple grade blocks are included allows the calculation of a kind of discriminant validity related to the positive effect of lower classroom sizes on academic performance.

Conclusion

The purpose of the literature review was to discuss both theoretical themes and empirical findings that are relevant to the problem identified in Chapter 1: the absence of insight into the effect sizes of class size on academic achievement. The theoretical discussion was based primarily on Abrantes et al.'s (2007) four-factor theory of classroom success with the review of empirical studies focusing on quantitative studies of the effect of class size on academic performance. The purpose of the conclusion of the literature review is to summarize and contextualize what is known of the relationship between classroom size and academic performance.

There appears to be consensus in the quantitative literature that there is a statistically significant and inverse correlation between classroom size and academic performance, such that students in smaller classes are likely to demonstrate improved academic performance. This finding has been explored and triangulated in a number of ways. In some instances, it has been demonstrated that moving students from larger to smaller classrooms improves the year-over-year performance of these students. For example, students who were in lower achievement percentiles when they were in larger classrooms have been observed to enter higher percentiles of achievement in smaller classrooms. It has also been observed that students who are otherwise similar in academic achievement levels and demographic qualities diverge in academic performance when randomly sorted into smaller and larger classes, with the students in smaller classes achieving higher levels of performance.

Studies that have tracked improvements in both same-student performance and similar-student performance as a response to the stimulus of changes in class size have taken a matched-pairs approach to demonstrating the positive effect of lower class sizes on academic performance. However, other statistical analyses have reached the same conclusion. For example, comparisons of students in smaller classes and larger classes have revealed that the students in smaller classes do better. Ultimately, the results of studies that compare students in smaller classes to students in larger classes are not as valid, in terms of isolating the effect of class size, as matched-pairs approaches, which have controlled for the impact of multiple factors (such as student demographics, prior academic achievement, and other factors) alongside class sizes as possible predictors of academic performance.

The fact that the same positive effect of lower class sizes on academic performance has been found in many studies, and through many different statistical approaches, represents a convincing form of triangulation (Cho et al., 2012; Glass, 1982; Hoxby, 2000; Matta et al., 2015; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980). There appears to be little doubt that there is, in fact, a positive effect of lower class sizes on academic performance (Cho et al., 2012; Glass, 1982; Hoxby, 2000; Matta et al., 2015; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980). The main point of contention in the literature is the question of why this effect insists.

Understanding the possible rationale for the existence of the positive effect of lower class sizes on academic performance is rendered more difficult by the existence of multiple terms for the same theoretical concepts. For example, teacher morale, motivation, engagement, and self-efficacy are part of the same theoretical cluster of concepts related to positive psychological effect (Fan & Williams, 2010; Klassen & Chiu, 2011; Usher & Pajares, 2009). These kinds of terms have been used loosely (in that they have been put forward as theoretical explanations without accompanying empirical testing relying upon validated psychometric scales) and interchangeably in many empirical studies, contributing to the lack of a consensus on why the positive effect of smaller class sizes on student achievement exists.

However, achieving a better understanding of the existence of the positive effect of lower class sizes on academic performance is a question of better empirical design as well as of more theoretical precision. Empirical research designs pertaining to the analysis of classroom size and academic performance can be improved by including the covariates of teacher experience, the percentage of special education students in a given class, the percentage of students of low socioeconomic status in a given class, and the number of minutes of paraprofessional or other co-teaching presence in a given class. Possible reasons for the existence of the positive effect of lower class sizes on academic performance can also be achieved by conducting studies in which multiple levels and grade blocks (such as elementary school, middle school, and high school) are included. The methodology and research design described and defended in Chapter 3 have been designed to close these observed gaps in the empirical literature.

Ultimately, even if the current study did not close existing gaps in the literature, it would still be worth carrying out because of the case-dependent dynamics of the link between classroom size and academic performance. The literature (Cho et al., 2012; Glass, 1982; Go & Lindert, 2010; Hanushek, 1999; Hoxby, 2000; Matta et al., 2015; Mosteller, 1995; Mueller, 2013; Rodriguez & Elbaum, 2014; Shapson et al., 1980) on this topic indicates that the nature of the link between classroom size and academic performance is likely to vary in different settings. Therefore, the existence of a positive link between smaller classroom size and improved academic performance in one environment does not mean that such a link exists in other environments. For this reason, there is value in carrying out an analysis of the link between classroom size and academic performance in any environment in which this link has not been studied before.

Chapter 3

Methodology

The purpose of this chapter is to describe and defend the various elements of methodology and research design. In order to do so, the chapter has been divided into distinct sections. First, statistical meta-analysis has been described and justified in the context of the research problem. Second, the research questions of the statistical meta-analysis have been presented. Third, the variables of the study are presented. Fourth, data analysis is explained. Finally, all of the methodological orientations of the meta-analysis have been summarized in a brief conclusion.

Description and Justification of Statistical Meta-Analysis

Based on the seminal literature (Campbell, 1988; Cochrane, 1972) on meta-analysis, meta-analysis is designed to achieve a reliability that is superior to that of isolated statistical analyses. In the standard approach to meta-analysis, effect sizes from different cases (which could include different published studies or different cases in a primary research analysis, such as different classrooms in a school) are calculated and then pooled together in order to provide an estimate of effect that is likely to be more reliable than the multiple estimates included as the input of the meta-analysis (Campbell, 1988; Cochrane, 1972). Statistical meta-analysis can be discussed with reference to the purpose of the study, which is to calculate the effect size of class size on student achievement in a single American school, Mohawk Area School District, in grades K through 8.

Statistically, there are several ways in which this purpose might be achieved. One approach is that of hierarchical linear modeling (HLM). One of the main pillars of HLM is nesting (Altman, 1991; Jackson, 2015; Kremelberg, 2010; Moore & McCabe, 2009; Natrella, 2013; Vogt & Johnson, 2011); for example, in an HLM analysis of class sizes and academic performance, students are nested within classes, classes within schools, schools within school districts, and school districts within regions or states. Given this nesting, researchers would have an interest in determining the effect of variation at any given level of the hierarchy on student achievement. For example, it might be that the link between class size and academic achievement varies from school to school; in one school, there might be an inverse correlation between class size and student achievement, while, at another school, there might be no significant correlation.

Within the context of the current study, an appropriate HLM structure would be as follows, with each arrow representing a level of nesting in the model.

Students→Classes→Grades

Because the study is taking place in the context of a single school, there are few nesting levels. Indeed, because the variable of class size can only be operationalized if the unit of analysis is the class, not the student, the nested structure of the study can be reduced every further, to:

Classes→Grades

Given that there is one nested level above that of class, the use of an advanced technique such as HLM for the current study appears to be unnecessary. Had the study included further nested levels—for example, proceeding up to the school district level—an HLM approach might have been more appropriate. However, given the relative simplicity of the structure of the study, a simple statistical approach, that of an effect size was chosen instead.

In this study, meta-analysis has been applied within the context of a case study. According to Yin (2009), a case study focuses on a research phenomenon "in depth and within its real-life context" (Yin, 2009, p. 18), and case studies are held to be appropriate when "the boundaries between phenomenon and context are not clearly evident" (p. 18). In the case of this study, the phenomenon of class size-student achievement interaction and the context of a single school are closely related. Because no other school's data has been included, the results of the study cannot cast light on the phenomenon of class size-student achievement interaction beyond the context of a single school, making this study a case study. In addition, the choice of a single research site means that the analysis will be based in a real-life context and will also be in depth, thus meeting Yin's other criteria for a case study.

Accordingly, the use of statistical meta-analysis for the study is justified on two grounds. First, there are not enough nested levels in the data structure to justify the use of an HLM approach, which might otherwise be considered appropriate. Second, statistical meta-analysis can generate a simple effect size calculation that can answer the research questions of the study. One way in which to understand the unique methodological contribution made by statistical meta-analysis is through the distinction between information and data provided by Dixon:

> Information [is] data that is 'in formation'—that, data that has been sorted, analyzed, and displayed, and is communicated through spoken language, graphic displays, or numeric tables. Knowledge, by contrast, is defined as the meaningful links people make in their minds between information and its application in action in a specific setting (Dixon, 2000, p. 13).

The school that is serving as the research site for the current study possesses information about student achievement (collected through i-Ready assessment tracking software) and class sizes, but lacks what Dixon described as knowledge about (a) the significance and magnitude, as measured through effect size, of the link between classroom size and student achievement; and (b) what kinds of classroom sizes should be sought by the school, given the objective of improving the academic performance of its students. The section on data analysis provided later in this chapter explains how the techniques of statistical meta-analysis can be utilized to generate knowledge from the school's existing information.

Research Questions

The measures of academic success in the study are reading performance and math performance. These indicators are being measured at three levels: (a) within each individual grade; (b) within the elementary and junior high school categories; and (c) for the school as a whole. Thus, while the research questions of the study can be expressed simply, in terms of the effect size measured when looking at the impact classroom size has on school-wide academic achievement, it is necessary to specify several sub-questions in order to ensure that the effects of classroom size on academic achievement within (a) individual grades; and (b) the elementary and junior high school categories are also captured.

RQ1: What is the effect size of classroom size on reading achievement at the school level?

RQ2: What is the effect size of classroom size on math achievement at the school level?

Tables 2 and 3 below contain the sub-research questions of the study. Cumulatively, these sub-research questions are designed to measure the effects of classroom size on academic achievement within (a) individual grades; and (b) the elementary and junior high school categories.

Table 2

RQ 1 and Sub-RQs of the Study

Research Question	Sub-Research Questions
RQ1: What is the effect size of classroom size on reading achievement at the school level?	1.1 What is the effect size of classroom size on reading achievement in grade K?
	1.2 What is the effect size of classroom size on reading achievement in grade 1?
	1.3 What is the effect size of classroom size on reading achievement in grade 2?
	1.4 What is the effect size of classroom size on reading achievement in grade 3?
	1.5 What is the effect size of classroom size on reading achievement in grade 4?
	1.6 What is the effect size of classroom size on reading achievement in grade 5?
	1.7 What is the effect size of classroom size on reading achievement in grade 6?
	1.8 What is the effect size of classroom size on reading achievement in grade 7?
	1.9 What is the effect size of classroom size on reading achievement in grade 8?
	1.10 What is the effect size of classroom size on reading achievement in grades K-6?
	1.11 What is the effect size of classroom size on reading achievement in grades 7-8?

Table 3

RQ 2 and Sub-RQs of the Study

Research Question	Sub-Research Questions
RQ2: What is the effect size of classroom size on math achievement at the school level?	2.1 What is the effect size of classroom size on math achievement in grade K? 2.2 What is the effect size of classroom size on math achievement in grade 1? 2.3 What is the effect size of classroom size on math achievement in grade 2? 2.4 What is the effect size of classroom size on math achievement in grade 3? 2.5 What is the effect size of classroom size on math achievement in grade 4? 2.6 What is the effect size of classroom size on math achievement in grade 5? 2.7 What is the effect size of classroom size on math achievement in grade 6? 2.8 What is the effect size of classroom size on math achievement in grade 7? 2.9 What is the effect size of classroom size on math achievement in grade 8? 2.10 What is the effect size of classroom size on math achievement in grades K-6? 2.11 What is the effect size of classroom size on math achievement in grades 7-8?

Study Variables

The research questions and sub-research questions of the study will be answered through the calculation of effect sizes using Cohen's *d*. The results of the Cohen's *d* on which the results will be based do not consist solely of the independent variable of class size and the dependent variable of academic achievement. Several control variables have also been included in the study. Table 4 contains all of the study variables.

Table 4

Variables in Study

Independent Variable	Control Variables	Dependent Variables
Class size	Teacher's years of experience	Reading performance
	Teacher's gender	Math performance
	% of students in special education	
	% of students of low socioeconomic (SES) status	
	Paraprofessional's minutes per day in classroom	

These variables will be utilized in the effect size calculations. The inclusion of these variables poses some challenges for effect size calculation. These challenges, and an approach to solving them, will be discussed in the subsequent section on data analysis.

Measurement

Of the variables listed in Table 4 above, only the variables of reading and math performance represent constructs rather than objectively measurable facets of reality. Therefore, particular attention has to be paid to how academic performance is conceptually and operationally defined in the study. Decisions related to the measurement of academic achievement should also be grounded in the existing theories and empirical approaches related to class size and academic achievement. In addition, because of the choice to represent academic performance longitudinally, the variable of class size also needs a more detailed operational definition.

In Nye et al.'s (2000) seminal analysis of the STAR program in Tennessee, one of the main theoretical assumptions was that the effect of small class sizes was cumulative. As discussed in the literature review, Nye et al. found that the achievement level of students going from small to small classes was greater than the achievement level of students going from small to medium or from small to large classes. The proposed theoretical explanation for this empirical finding was that students who spend time in small classes improved over time, probably due to exposure to some variation of the academic improvement variables discussed by Abrantes et al. (2007). If it is indeed the case that exposure to smaller class sizes has longitudinal or somehow cumulative effects on student achievement, then a particularly appropriate measure of performance would be relative improvement rather than absolute performance.

The advantage of defining academic achievement in terms of performance is that such an approach reduces bias and maximizes comparability across classrooms. It is possible that some students are sorted in classrooms on the basis of factors that cannot be statistically controlled for; for example, there could be a flaw in an automated system that leads to the assignment of the best students to smaller classes. In such a case, the finding of an effect of class size on academic achievement would be spurious, as this finding would be causally attributable not to class size itself, but rather to the quality of the students in the classroom, regardless of size. In any scenario in which there is not a truly random assignment of students to classrooms, it remains a risk that some third variable, such as student skill level, will exert a hidden influence on the data and thus render the study's results less valid and reliable.

Treating academic performance as improvement removes this methodological obstacle. In the i-Ready system used by the school, student performance data are available throughout the school year. This means of operationalizing the academic performance variable reasons the chances that the effect size measurements proposed in the data analysis section will accurately identify the effect of class sizes themselves.

Data Collection

Data collection for the study will be facilitated by the school that is serving as the research site. The school will provide data for each class in the school, encompassing all of the variables listed in Table 4. These data will be depersonalized in order to protect the identity of the teachers. Because the performance data are aggregate measures, they protect both the privacy and anonymity of students, as the data cannot be used to identify any students. The data will be provided into Microsoft Excel format and will draw upon assessment reports from i-Ready (for the academic performance scores) as well as from the school's own internal reporting system (for the other variables listed in Table 4).

Reliability and Validity

The difference between reliability and validity is that reliability is "the degree to which measures are free from error and therefore yield consistent results" (Zikmund, 2003, p. 302) whereas validity is "the ability of a measure to measure what it is supposed to measure" (Zikmund, 2003, p. 302). Both reliability and validity are important concerns in any quantitative study (Balnaves & Caputi, 2012; Creswell, 2015; Creswell & Plano Clark, 2011; Davies & Hughes, 2014; Jackson, 2015; Leary, 2011; McBurney & White, 2011; McNabb, 2010; Trochim et al., 2015; Yin, 2009; Zikmund, 2003). Accordingly, this section of the methodology contains a discussion of the reliability and validity risks and mitigation strategies in the study.

One of the main threats to the reliability of the study is that data from a single year could reflect the effects of chance more than the systematic effect of class sizes on academic achievement. It is possible that, in a single year, there was a negative correlation between classroom size and academic achievement, but that this correlation was positive or did not exist in other years. Thus, choosing only a single year for analysis constitutes a threat to the reliability of the study. The proposed mitigation strategy for this threat to reliability is to conduct what statisticians (Altman, 1991; Jackson, 2015; Kremelberg, 2010; Moore & McCabe, 2009; Natrella, 2013; Vogt & Johnson, 2011) have referred to as robustness tests, meaning attempts to determine whether a particular finding remains after it has been investigated in more than one way. In this study, the robustness test chosen to ensure the reliability of the measured effect of class sizes on academic achievement is testing on multiple years. An analysis was conducted over multiple school years: 2015-2016 and 2016-2017.

The validity of the study is based on the premise that the academic performance data actually measure academic performance. The use of i-Ready, whose assessments are based on existing standards, is one possible guarantee of the validity of the study. The i-Ready Diagnostic is backed by extensive validity research and consistently achieves results. Independent research showed i-Ready Diagnostic to be highly correlated with Common Core state assessments. However, one threat to the validity of the study can be found in the definition of class size. For the purposes of this study, class size has been coded as a continuous variable. However, it is not necessarily the case that the effects of class size can best be measured by treating class size as a continuous variable. In a seminal study of the effects of the STAR classroom size pseudo-experiment, Nye et al. (2000) operationalized classroom size polytomously, differentiating between small, medium, and large classrooms. Such an approach is not only methodologically pragmatic but also statistically useful.

Nye et al.'s findings suggest that the possible effects of classroom size might be found not in minute changes, but rather in big ones. To offer a practical illustration of the logic behind Nye et al.'s polytomous coding scheme, it might not be the case that a transition from an 22-student to a 19- or 20-student classroom would have an impact on academic achievement, but it might be the case that transitioning from a 22-student classroom to an 8-student classroom would have a substantial impact. What Nye et al. utilized was an analysis of variance (ANOVA) approach in which there were three predefined levels (small, medium, and large), and the means of achievement were measured and compared at each of these levels. The advantage of the ANOVA approach taken by Nye et al. is that it can capture differences between broad categories, in particular, small, medium, and large classrooms.

Conclusion

The purpose of this chapter was to describe and defend the various elements of methodology and research design. First, statistical meta-analysis was justified, with a method of calculating effect sizes using Cohen's *d* over the alternate HLM method. Second, the research questions and sub-research questions of the study were presented. Third, the variables of the study—including the independent variable of class size; the control variables of teacher's years of experience, teacher's gender, percentage of students in special education, percentage of students of low socioeconomic (SES) standing, and the paraprofessional's minutes per day in the classroom; and the dependent variables of reading performance and math performance—were presented. Fourth, data analysis was explained, with an emphasis on how the independent effect size of class sizes can be calculated.

Chapter 4

Findings

Introduction

The primary purpose of this individual participant meta-analytic investigation was to examine the effect of class size on reading and math achievement based on the i-Ready scaled scores. This quantitative case study is based on statistical meta-analysis to calculate the effect size of class size on student achievement in the Mohawk Area School District. This purpose was achieved through calculating the effect size using Cohen's *d*. Cohen (2013), an authority on effect sizes, has stated that an effect size of 0.2 is small, an effect size if 0.5 is medium, and an effect size of 0.8 is large. In the standard approach to meta-analysis, effect sizes from different cases (which could include different published studies or different cases in a primary research analysis, such as different classrooms in a school) are calculated and then pooled together in order to provide an estimate of effect that is likely to be more reliable than the multiple estimates included as the input of the meta-analysis (Campbell, 1988; Cochrane, 1972).

This meta-analytical investigation includes a pooling of effect sizes from multiple classrooms into single effect size figures. This investigation also took into account an analysis at each grade level, K-8, and looked at several variables; teacher's years of experience, teacher's gender, the percentage of students in special education through an Individualized Education Plan (IEP), the percentage of students from a low socioeconomic status (SES) as identified by the free/reduced lunch data, and the number of minutes that a paraprofessional was assisting in a classroom for a daily period. As indicated in Chapter 3, teacher's years of experience, teacher gender, and paraprofessional minutes per day in the classroom were to be included variables and considered as part of this meta-analytic study. Once the final collection of the data was gathered, it was decided to not include these variables for various reasons. The teacher's years of experience was not reported because it would be easy to determine the teacher name based on the limited number of first year teachers and the few teachers near the end of their career because the study was conducted in one district, the Mohawk Area School District. The teacher gender in grades K through eight is weighted heavily by females and did not have enough males to make that delineation valid. The paraprofessional minutes per day in the classroom were not varied enough because most classrooms had the same number of minutes of paraprofessional support. Therefore, the percentage of students having an IEP in a particular classroom and the percentage of students of low SES were included in this study because they were valid measures to examine.

The chapter is organized by each research question. Therefore the results related to RQ1 are presented, followed by the findings for RQ2. The results germane to the sub-research questions are embedded throughout the chapter. A brief conclusion summarizes the main findings of the study.

Research Question 1 Results

The first research question of the study was as follows: What is the effect size of classroom size on reading achievement at the school level? This question was answered by calculating d for classes of 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, and 27 students as defined in Appendix B.

For overall reading achievement:

$$d(97) = -1.266, \; p<.001$$

This effect size indicates that there is significant variability across the reading scores. There is no pattern to the effect size as depicted in Table 5 as it moves from smaller to larger class sizes as it relates to reading achievement.

Table 5

Reading Achievement by Class Size

CLASS SIZE	NUMBER OF STUDIES	EFFECT SIZE
12	1	0.119
14	2	0.579
15	5	0.229
16	7	0.322
17	5	1.003
18	8	-1.014
19	10	-1.289
20	17	0.277
21	7	0.815
22	10	1.146
23	16	1.062
24	2	0.553
25	4	1.701
26	2	1.363
27	1	0.960

Interestingly, the two reading classes on the opposite ends of the class size spectrum represent two different bins of class size; the class of n = 12 students represents smaller class size, whereas the class of n = 27 students represents a larger class size. The fact that both small and large classes have a similar effect size, strongly suggests that, in reading, there is no meaningful effect of class size on performance.

To answer the sub-research questions, the effect size across grade levels was examined; the analysis indicating that significant heterogeneity exists within and between each grade level. At the seventh and eighth grade level, the most positive effect occurs which might reflect an artifact of the i-Ready assessment diagnostic capability. The scaled score increases as you go through the grade levels from kindergarten through eighth grade of the diagnostic assessment.

The grade level analysis follows the expected pattern with the exception of grade six which does not progressively increase, similar to the other grade levels. The following table summarizes class size effect by grade level (specific data in Appendix C) on i-Ready achievement in reading where $p<.001$.

Table 6

Reading Achievement by Grade Level

GRADE LEVEL	NUMBER OF STUDIES	EFFECT SIZE
Kindergarten	12	-1.289
First Grade	11	-0.561
Second Grade	11	0.360
Third Grade	11	0.354
Fourth Grade	11	1.002
Fifth Grade	10	1.142
Sixth Grade	9	0.562
Seventh Grade	11	1.562
Eighth Grade	11	1.562

Next, the potential moderators Percentage of Special Education Students and Percentage of Low SES in a classroom are examined. Table 7 (detailed data in Appendix D) provides the mean effect size measures, based on the proportion of special education students present in a class.

Table 7

Percentage of Students in Special Education in the Reading Classroom

PERCENT OF STUDENTS IN SPECIAL EDUCATION	NUMBER OF STUDIES	EFFECT SIZE
0%-10%	61	0.622
11%-20%	25	0.952
21%-30%	7	-1.289
31%-40%	4	0.075

Results indicate that the strongest positive effect is found when there are 11-20% students in classes identified as special education. Special education instructional paraprofessional support is consistent in all classrooms. Those students would receive additional time for individualized instructional assistance because of the fewer number of special education students in that particular classroom.

Table 8 provides the results for Percentage of Low SES Students by reading achievement; a detailed description of the data is provided in Appendix E.

Table 8

Percentage of Students of Low Socioeconomic Status (SES) in the Reading Classroom

PERCENT OF STUDENTS OF LOW SOCIOECONOMIC STATUS	NUMBER OF STUDIES	EFFECT SIZE
0%-10%	1	1.810
11%-20%	5	1.117
21%-30%	14	1.215
31%-40%	25	0.199
41%-50%	28	-1.288
51%-60%	15	0.947
61%-70%	9	0.569

These results present an expected pattern with the exception of the 41%-50% group. This group contains the largest number of classes (28) but this is not a significant number of studies greater than that of the 31%-40% group which contains n = 25 studies. Extant research suggests that as the percentage of students in the low SES group increases the level of reading achievement would decrease (Burkam & Lee, 2002). The more positive effect of the 51%-60% range and the 61%-70% range may indicate an increase in paraprofessional support in those particular classrooms through the Title 1 or special education programming. The elementary building (Grades K-6) is a school-wide Title 1 building so classrooms are supported by the Title 1 reading specialist or an instructional paraprofessional.

Research Question 2 Results

The second research question of the study was as follows: What is the effect size of classroom size on math achievement at the school level? This question was answered by calculating *d* for classes of 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, and 27 students; a detailed account of the data is in Appendix F.

For overall math achievement:

$$d(93) = -0.671, \quad p<.001$$

which indicates that there is significant variability across the math scores. There is no pattern to the effect size as depicted in Table 9 as it moves from smaller to larger class sizes as it relates to math achievement. What is interesting about the effect of class size on math achievement, there are only two class sizes with a positive effect size; classes of n = 25 and n = 27 students.

Table 9

Math Achievement by Class Size

CLASS SIZE	NUMBER OF STUDIES	EFFECT SIZE
12	1	-0.587
13	2	-0.333
14	3	-0.349
16	6	-1.209
17	3	-0.371
18	11	-0.808
19	8	-1.273
20	11	-1.052
21	8	-0.276
22	8	-0.388
23	16	-0.558
24	10	-0.123
25	3	0.002
26	2	-0.189
27	1	0.101

Examining the math achievement effect across grade levels to answer the sub-research questions, the analysis indicates that within each grade level there is significant heterogeneity. Across all the groups significant differences also exist as it did with the analysis of the reading achievement data. At the seventh and eighth grade level the most positive effect occurs which is most likely an artifact of the i-Ready assessment diagnostic capability. The scaled score increases as you go through the levels of the diagnostic assessment just as it does with reading achievement.

The grade level analysis follows the expected pattern of progressively increasing as you go through each grade level. Table 10 summarizes class size effect by grade level on i-Ready achievement in math where $p<.001$. The data details for each grade level are documented in Appendix G.

Table 10

Math Achievement by Grade Level

GRADE LEVEL	NUMBER OF STUDIES	EFFECT SIZE
Kindergarten	12	-1.705
First Grade	11	-1.213
Second Grade	11	-0.832
Third Grade	11	-0.559
Fourth Grade	8	-0.399
Fifth Grade	10	-0.215
Sixth Grade	9	-0.361
Seventh Grade	10	0.099
Eighth Grade	11	0.106

Examination and analysis of the math i-Ready achievement results by other variables are indicated in the tables below. Next, the potential moderators Percentage of Special Education Students and Percentage of Low SES in a classroom are examined with the details of the data depicted in Appendix H. Table 11 provides the mean effect size measures, based on the proportion of special education students present in a class.

Table 11

Percentage of Students in Special Education in the Mathematics Classroom

PERCENT OF STUDENTS IN SPECIAL EDUCATION	NUMBER OF STUDIES	EFFECT SIZE
0%-10%	57	-1.142
11%-20%	21	-0.559
21%-30%	9	-1.046
31%-40%	6	-0.364

Results indicate that the most positive effect (although negative) is found when there are 31-40% students in classes identified as special education. The support in these classrooms is conducted by a special education teacher and/or an instructional paraprofessional. The most positive effect may be indicative of a stronger support system with more minutes focused on individualized instructional support. Table 12 provides the results for Percentage of Low SES Students by math achievement (see Appendix I for the data specifics).

Table 12

Percentage of Students of Low Socioeconomic Status (SES) in the Mathematics Classroom

PERCENT OF STUDENTS OF LOW SOCIOECONOMIC STATUS	NUMBER OF STUDIES	EFFECT SIZE
11%-20%	4	-1.233
21%-30%	14	0.051
31%-40%	22	-0.550
41%-50%	29	-0.740
51%-60%	16	-1.049
61%-70%	8	-0.991

Math achievement by low SES does not follow the expected pattern as is depicted in Table 12. The lowest SES group (11%-20%) contains the largest negative effect. Extant research suggest that it is reasonable to expect that as the percentage of students in the low SES group increases the level of math achievement would decrease (Burkam & Lee, 2002). This finding might be due to the small sample size; it is supported by only four classrooms. Supports in the classroom through Title 1 and special education may be indicative of the 14 classrooms with 21%-30% low SES.

Table 13 is a comparison between the math and reading achievement results as related to class size. Class size impacts reading achievement more than math achievement based on the data detailed in Appendix J. These differences are statistically significant ($p < .001$).

Table 13

Reading Achievement and Math Achievement Effect Size

SUBJECT AREA	NUMBER OF STUDIES	EFFECT SIZE
READING	97	-1.266
MATH	93	-0.671

Regression Analysis

In an effort to provide a pragmatic model for use within the Mohawk Area School district, a meta-analytic regression model was created to address both RQ1 and RQ2. These analyses were conducted on both reading and math scores, examining the class size moderator, the grade level moderator, and interaction of class size and grade level. The interaction effect was computed by taking the product of class size and grade level (Tabachnick & Fidell, 2013). The interaction effect was not significant so it was removed from the model in both reading and math.

In the running regression for reading, $R^2 = 0.78$ which indicates that the equation below will fit with Mohawk's model of i-Ready reading assessment 78% of the time. Statistically speaking, it is a good predictor of what a reading classroom mean i-Ready reading score should be during the spring assessment. The reading data is described in Appendix K.

$Y_i = -1.9188 + (\text{class size})(.0519) + (\text{grade level})(.3541)$ is the equation to use for the effect size. An equation that is practical for classroom teacher use contains a regression model using true data:

$$Y = 365.905 + (\text{class size})(3.786) + (\text{grade level})(26.161)$$

based on a standard regression model:

$$Y_i = b_0 + b_1X_1 + b_2X_2 + e_1$$

This model indicates that for student "i", the expected reading score will be equal the slope intercept (b_0), plus the class size times the slope for class size(b_1X_1), the grade level times the slope for grade level(b_2X_2), and error associated with student i (e_1). Based on the R^2, it is expected that this model will average 78% accuracy in predicting students reading scores. This determination will indicate to a teacher what their average score for their classroom should be to follow Mohawk's predictable past pattern of i-Ready achievement scores over the past two years. Of course, the goal would be to surpass the average score to ensure continued exceptional growth within one's classroom.

Figure 1 provides a graphical depiction of the relationship between the mean effect size and the class size for reading in a scatterplot.

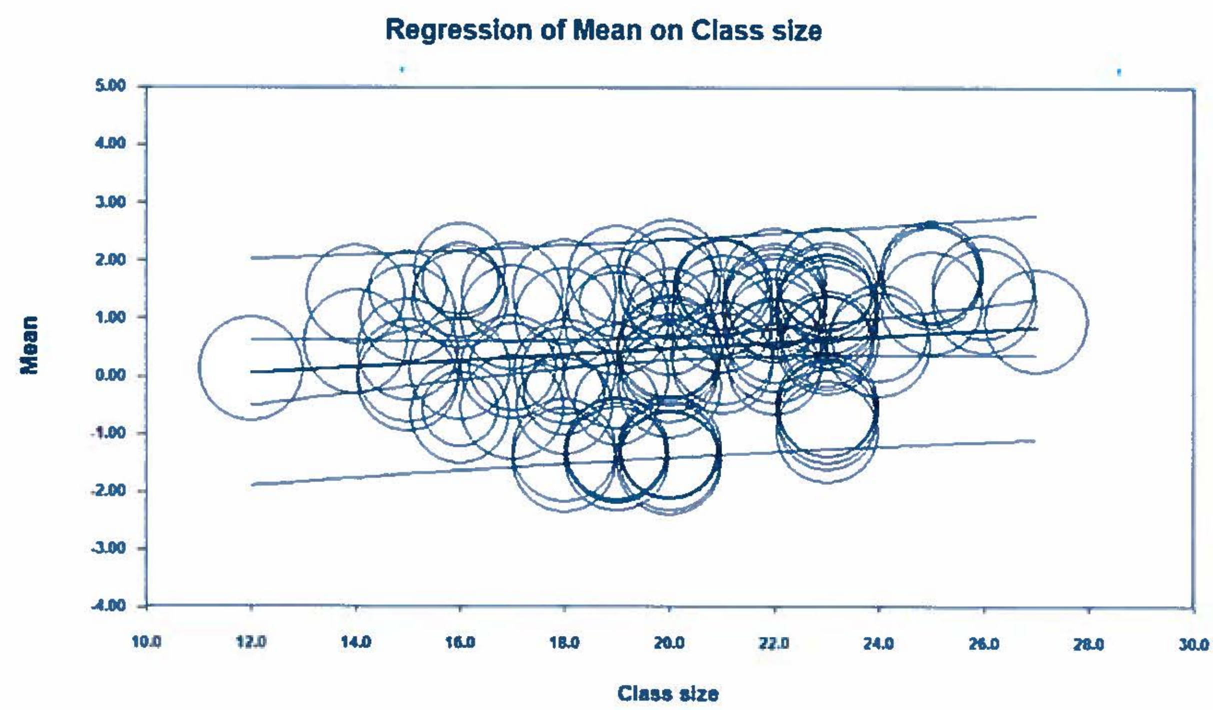

Figure 1. *Scatterplot Representation for Reading-Class Size*

Figure 2 shows the relationship between the mean effect size and the grade level for reading.

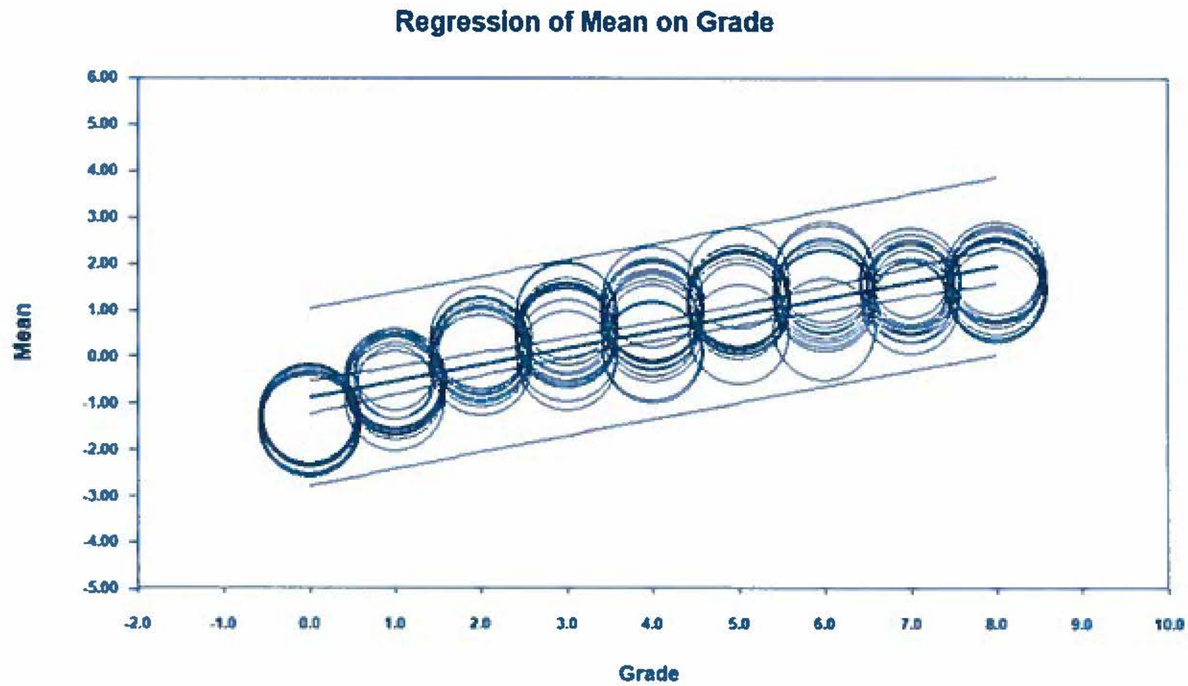

Figure 2. *Scatterplot Representation for Reading-Grade Level*

The same model estimation was conducted with math achievement. The results of the math regression model indicated $R^2 = 0.82$ which indicates that the equation below will fit with Mohawk's model of i-Ready math assessment 82% of the time. The data is detailed in Appendix L. Similar to the analysis conducted for Reading scores, the model for the Math scores was found to be: $Y_i = -2.0467 + (\text{class size})(.0307) + (\text{grade level})(.2198)$ is the equation to use for the effect size. Likewise, an equation using true data:

$$Y = 352.208 + (\text{class size})(2.48) + (\text{grade level})(16.554)$$

based on a standard regression model:

$$Y_i = b_0 + b_1X_1 + b_2X_2 + e_1$$

This model indicates that for student "i", the expected math score will be equal the slope intercept (b_0), plus the class size times the slope for class size(b_1X_1), the grade level times the slope for grade level(b_2X_2), and error associated with student i (e_i). Based on the R^2, it is expected that this model will average 82% accuracy in predicting students' math scores. This determination will indicate to a teacher what their average score for their classroom should be to follow Mohawk's predictable past pattern of i-Ready math achievement scores over the past two years. The goal would be to surpass the average math achievement score to ensure continued exceptional growth within one's classroom.

Figure 3 shows the relationship between the mean effect size and the class size for math in a scatterplot.

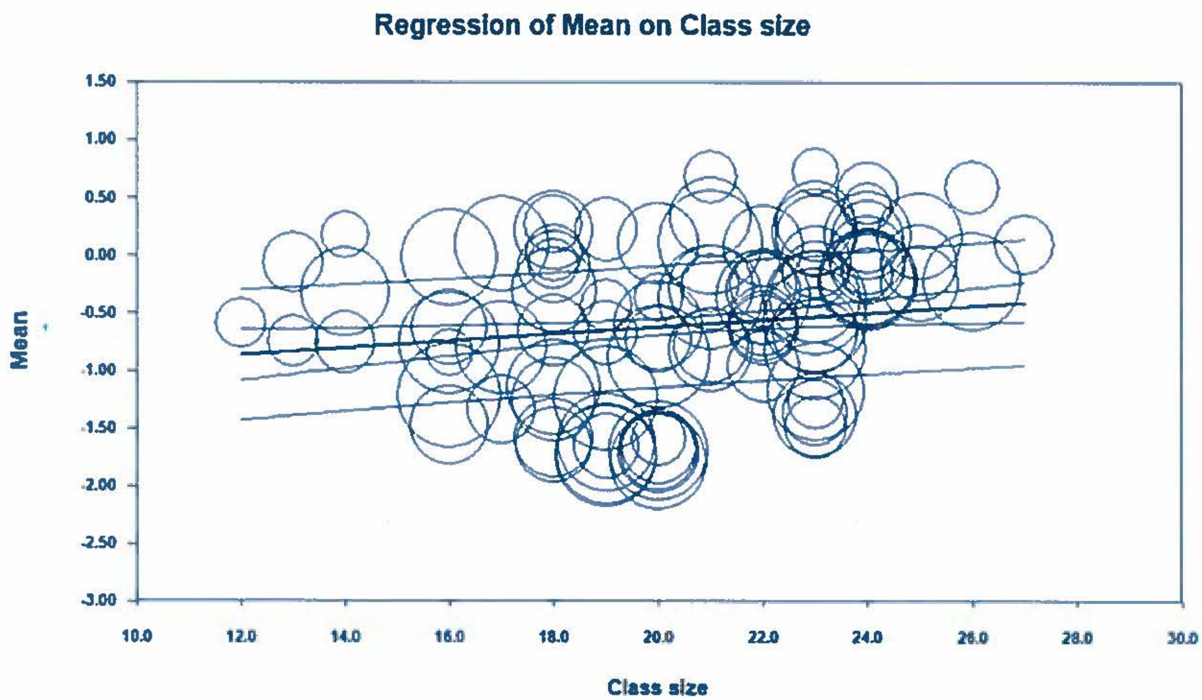

Figure 3. *Scatterplot Representation for Math-Class Size*

Figure 4 shows the relationship between the mean effect size and the grade level for math.

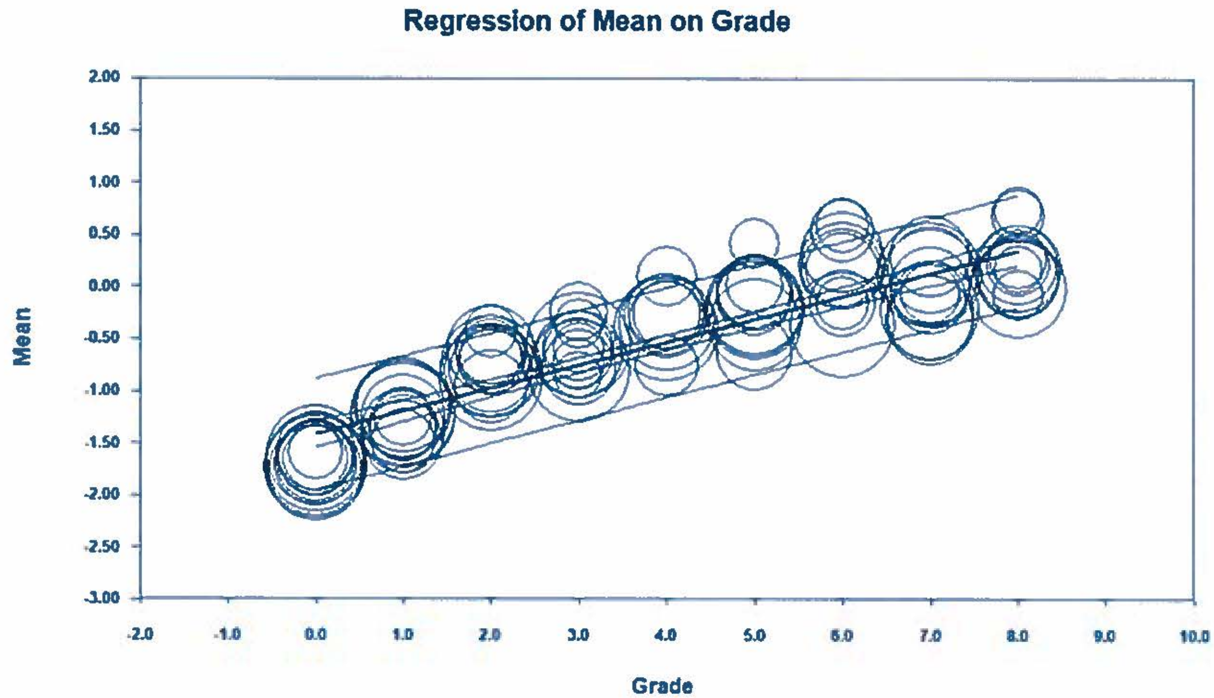

Figure 4. *Scatterplot Representation for Math-Grade Level*

Conclusion

The purpose of this quantitative, effect size-based study was to examine the relationship between class size and two kinds of achievement in the context of a single school district. The two kinds of achievement that were measured were reading achievement and math achievement. The study found an absence of meaningful effects. In terms of reading achievement, the effect size measure indicated heterogeneity across the class data. The effect size indicated the existence of significant variability across reading scores. The effect sizes for reading achievement gave no significant evidence that classes of any particular size (small, medium, or large) were more likely to be associated with higher levels of reading achievement.

In terms of math achievement, the effect size measure also indicated heterogeneity across the class data. As with reading, the *d* for math scores indicated significant variability across the math scores. In addition, as with reading, there was no significant evidence that classes of any particular size (small, medium, or large) were more likely to be associated with higher levels of math achievement.

Using regression models, it was found that each additional student in math classes improved performance by 0.0307 points, whereas each additional student in reading classes increased performance by 0.0519 points. These coefficients were very close to 0, further indicating the absence of a true effect of class size on reading or mathematics performance. Thus, the findings of the study did not confirm previous literature review findings about the significant effect of class sizes on student achievement in both reading and math, in which smaller classes as compared to larger class sizes were found to be associated with higher levels of performance.

CHAPTER 5

Conclusions and Recommendations

Introduction

As a benefit to the reader, this concluding chapter of the dissertation restates the research problem and reviews the method used in the study. The main purposes of this chapter of the study are to (a) review the findings, (b) discuss the findings of the study, (c) acknowledge the limitations of the study, (d) provide recommendations for further research, and (e) provide recommendations for practice.

Statement of the Problem

The issue of class size has been a topic in America since the 1950's. It wasn't until the 1970's that the impact of class size on academic achievement was examined in detailed studies. Educators have indicated that with smaller class sizes they are able to produce better student achievement results in reading and mathematics. With the increase in teacher accountability, class size is a topic discussed at the school district level.

The problem is that the relative lack of effect size calculations in the empirical literature does not provide policy-makers, educational leaders, and other relevant stakeholders with the kind of information needed to support decisions to reduce class sizes. In scholarly terms, the problem is rooted in the difference between statistical significance and effect sizes. It is possible for there to be a statistically significant relationship of small class sizes on student achievement, but for the actual magnitude of the effect of class sizes on student class sizes to be too small to warrant changes to class sizes. In practical terms, the problem is that decision-makers lack the kind of data support they need to be able to settle on appropriate class sizes or to set targets for class sizes.

The purpose of this quantitative case study based on statistical meta-analysis is to calculate the effect size of class size on student achievement in the Mohawk Area School District, Grades K-8. This purpose was achieved through calculating the effect size using Cohen's *d*. Measuring the relationship between classroom size and academic performance in the context of several classes increased the reliability of overall conclusions related to the relationship between classroom sizes and academic performance.

Review of the Methodology

The study was a quantitative case study based on the use of statistical meta-analysis. The study is quantitative because its main feature is the calculation of a mathematical parameter, an effect size, on the basis of variations in mathematically coded independent, control, and dependent variables. The study is a case study because it is limited to one school. Finally, the study is meta-analytical in that it includes a pooling of effect sizes from multiple classrooms into single effect size figures.

Meta-analysis is designed to achieve a reliability that is superior to that of isolated statistical analyses. In the standard approach to meta-analysis, effect sizes from different cases (which could include different published studies or different cases in a primary research analysis, such as different classrooms in a school) are calculated and then pooled together in order to provide an estimate of effect that is likely to be more reliable than the multiple estimates included as the input of the meta-analysis (Campbell, 1988; Cochrane, 1972). Statistical meta-analysis was used to calculate the effect size of class size on student achievement in a single American school, Mohawk Area School District, in grades K through 8. The research questions of the study were answered through the calculation of effect sizes using Cohen's *d*.

RQ1: What is the effect size of classroom size on reading achievement at the school level?

RQ2: What is the effect size of classroom size on math achievement at the school level?

The results of the Cohen's *d* on which the results will be based do not consist solely of the independent variable of class size and the dependent variable of academic achievement. Two other variables were analyzed in this study: the percentage of students in special education in each classroom and the percentage of students with low socioeconomic status based upon free/reduced lunch data.

The robustness test chosen to ensure the reliability of the measured effect of class sizes on academic achievement is testing on multiple years. This analysis was conducted over the 2015-2016 and 2016-2017 school years. The validity of the study was based on the premise that the academic performance data actually measures academic performance. The use of i-Ready, whose assessments are based on existing standards, was one possible guarantee of the validity of the study.

Summary of the Results

The first research question yielded results about the effect of class size on reading achievement. In terms of reading achievement, the effect size measure indicated heterogeneity across the class data. The effect size indicated the existence of significant variability across reading scores. The effect sizes for reading achievement gave no significant evidence that classes of any particular size (small, medium, or large) were more likely to be associated with higher levels of reading achievement.

Upon examination of the other variables in relationship to class size and reading achievement, the following results were observed.

- Results indicate that the strongest positive effect is found when there are 11-20% students in classes identified as special education. Special education instructional paraprofessional support is consistent in all classrooms. Those students would receive additional time for individualized instructional assistance because of the fewer number of special education students in that particular classroom.
- These results for the percentage of low SES present an expected pattern with the exception of the 41%-50% group. This group contains the largest number of classes (28) but this is not a significant number of studies greater than that of the 31%-40% group which contains n = 25 studies. Extant research suggests that as the percentage of students in the low SES group increases the level of reading achievement would decrease (Burkam & Lee, 2002). The more positive effect of the 51%-60% range and the 61%-70% range may indicate an increase in paraprofessional support in those particular classrooms through the Title 1 or special education programming. The elementary building (Grades K-6) is a school-wide Title 1 building so classrooms are supported by the Title 1 reading specialist or an instructional paraprofessional.

In terms of math achievement, the effect size measure also indicated heterogeneity across the class data. As with reading, the d for math scores indicated significant variability across the math scores. In addition, as with reading, there was no significant evidence that classes of any particular size (small, medium, or large) were more likely to be associated with higher levels of math achievement.

Upon examination of the other variables in relationship to class size and math achievement, the following results were observed.

- Results indicate that the most positive effect (although negative) is found when there are 31-40% students in classes identified as special education. The support in these classrooms is conducted by a special education teacher and/or an instructional paraprofessional. The most positive effect may be indicative of a stronger support system with more minutes focused on individualized instructional support.
- Math achievement by low SES does not follow the expected pattern. The lowest SES group (11%-20%) contains the largest negative effect. This finding might be due to the small sample size; it is supported by only four classrooms. Supports in the classroom through Title 1 and special education may be indicative of the 14 classrooms with the only positive effect of 0.051 which was indicative of the 21%-30% low SES group that was examined.

The most beneficial result of the study for the Mohawk Area School District is the meta-analytic regression model for both reading and math i-Ready assessment data.

Discussion of the Results

The main finding of the study was that class size was not significantly associated with performance in either reading or mathematics. This finding was unexpected. Based on a review of the previous literature (Cho, Glewwe, & Whitler, 2012; Glass, 1982; Goldstein & Blatchford, 1998; Hanushek, 1999; Hattie, 2005; Hoxby, 2000; Mosteller, 1995; Mueller, 2013; Nye, Hedges, & Konstantopoulos, 2000; Shapson, Wright, Eason, & Fitzgerald, 1980; Watson, Handal, Maher, & McGinty, 2013), as presented in the second chapter of the study, it was expected that student performance would be higher in smaller classes and lower in larger classes. The hypothesized theoretical reason for such a finding was that, in larger classes, students would be less able to receive and benefit from individualized instruction from the teacher (Adkins-

Coleman, 2010; Houchen, 2013; Ross, Bondy, Gallingane, & Hambacher, 2008). Another possible reason for the positive effect of lower class sizes on academic achievement might be that, in smaller classes, students might have to interact more frequently, both with teachers and in the context of group work (Blondal & Adalbjarnardottir, 2012; Bobbitt-Zeher, 2004; Fan & Williams, 2010; Sun, 2014).

Both of these proposed mechanisms for the effectiveness of smaller classes in terms of student achievement rely on the construct of student engagement. If students are sorted into classes of different sizes randomly—as is highly likely in the case of a single school or school district—then one reason that low class sizes might be affected in terms of academic achievement is that of engagement, which could occur in the form of (a) teacher's customizing teaching to individual students or small groups of students or (b) students interacting more routinely with teachers and with other students. One possible interpretation of the findings from chapter 4 is that engagement is no greater in small classes than it might be in large or medium-sized classes.

Another possible interpretation is that the true effect of class sizes in engagement was not measured in the study, as the classes were neither large nor small enough to allow engagement to be measured. The sample size for small classes and large classes was minimal in comparison to the number of sample sizes from the medium class group. This could be a possible rationale for the discrepancy between the literature review and the actual meta-analytic study conducted within the Mohawk Area School District.

On the assumption that the effect size findings of the study are valid and reliable, the interpretative possibilities are that: (a) The previous empirical findings identifying a positive impact of lower class size are mistaken or (b) the same explanatory dynamics that applied in previous studies do not apply to the school district that provided the data for the current study. The second possibility is more likely. Many of the previous studies on the effect of classroom size on student performance are based on the aggregation of tens, hundreds, or thousands of student results. Because the current study contained information from a smaller sample of students, its results are highly unlikely to be generalizable to the population of all students.

Nonetheless, the local non-significance of the findings still offers important empirical and theoretical insights. It is certainly possible that the observed positive effect of smaller class sizes on academic achievement in reading and math is a function not of class sizes themselves, but of hidden variables. For example, it could be the case that schools with smaller class sizes also happen to be schools in wealthier school districts that are able to afford better curricular materials and otherwise improve the academic outcomes of their students. Only in the Tennessee STAR study was a conscious attempt made, across an entire state, to randomize class size assignment in order to measure the true effect of class size on student achievement, but these results are now dated and might not reflect current realities related to the connection between class size and student achievement. Few other class size studies have drawn on truly randomized data, and so it remains possible that, instead of measuring class size effects, they are measuring the effects of some third variable.

In the current study, the third-variable problem was mitigated by examining the same school district, which, because of its homogeneity, is unlikely to contain third-variable effects. The absence of an impact of classroom size on student achievement could nonetheless reflect some difference between the school district and other school districts. It could be, for example, that the school district possesses teachers with a higher level of experience or professional development. The absence of an effect of class size can, therefore, be considered a strength, because, in this school district, teachers are able to facilitate academic achievement even in larger classes, which is not the case in the previous classroom effect size studies. The non-significance of the effect of classroom size on academic achievement is therefore an opportunity for the school district to recognize, and, subsequently, better understand the reasons for its own success. The written curriculum in reading and math for the Mohawk Area School District may be implemented with fidelity across grade levels and across varying classes with different sizes. The faculty utilizes the same resources across each classroom in a grade level and the teaching professionals are able to ensure an increase in student achievement no matter how many students are included in their classrooms.

The third-variable problem was also actively addressed in the study through the inclusion of variables such as socioeconomic status and special education. These kinds of variables were found to have effects of their own, effects that have also been identified in prior literature. The fact that socioeconomic status and the density of special education students influenced student performance, but that class size did not, confirmed that the results obtained for the school district are unusual. Further study of how and why the school might have been able to surmount the previously observed negative impact of class size on student performance is warranted.

The pragmatic model derived from the study specifically for the Mohawk Area School District is of great benefit. This regression model with a 78% accuracy for reading and an 82% accuracy for math will indicate to a teacher what their average score for their classroom should be to follow Mohawk's predictable past pattern of i-Ready achievement scores over the past two years. The goal would be to surpass the average score to ensure continued exceptional growth within one's classroom.

The major limitation of this study is that it was conducted with one school district but the benefit from the study is significant and meaningful to the Mohawk Area School District. The study was also limited in that the details about teachers that were collected were not utilized because of the fact that only one district was included in the study. Teacher anonymity was important and if results were reported utilizing teacher specifics their identity could have been compromised. The individual properties of teachers are likely to influence their ability to manage larger classrooms in a manner that facilitates learning. It may have been useful to measure and include predictor variables such as teacher age and number of years of service in the statistical model of the study but due to respecting teacher anonymity; these factors were not included.

Recommendations for Future Research

A class-based effect size approach, although the approach taken in the current study, is not the only possible means of explaining the relationship between class size and student performance. In this study, each class's effect was measured separately based on the number of students in that class. It is also possible to aggregate the data so that the effect size is calculated not on the basis of individual classes but on the basis of classes that are grouped into categorical containers such as small, medium, and large classes. For example, instead of calculating class size effects separately for classes of 10, 12, 14, 18, 20, 21, 24, and 29 students, such an approach could involve grouping the classes of 10, 12, and 14 into a small category, the classes of 18, 20, and 21 into a medium category, and 24 and 29 into a large category.

Another interesting topic to be considered for future research is to determine what average i-Ready score is needed for a classroom to be proficient as a whole on the state standardized assessments (PSSA). The relationship between PSSA proficiency and a specific i-

Ready score would have to be determined and then it could calculated along with class size and grade level to create a regression model for this purpose. This would be very practical information for a classroom teacher.

Recommendations for Practice

The research literature supports the conclusion that, all else being equal, smaller classes are better (Cho et al., 2012; Glass, 1982; Goldstein & Blatchford, 1998; Hanushek, 1999; Hattie, 2005; Hoxby, 2000; Mosteller, 1995; Mueller, 2013; Nye et al., 2000; Shapson et al., 1980; Watson et al., 2013). However, it might not always be possible for schools or school districts to ensure that class sizes are small. In some cases, districts might be overcrowded and underfunded, resulting in a lack of ability to keep class sizes small. Nonetheless, whenever it is possible to keep class sizes small, it seems more likely that teachers and students will engage with each other better, which should lead to better academic outcomes. In the current study, a significant relationship between classroom size and student achievement was not found. However, this finding should not be taken to support the recommendation that schools allow class sizes to swell. Based on the previous findings (Cho et al., 2012; Glass, 1982; Goldstein & Blatchford, 1998; Hanushek, 1999; Hattie, 2005; Hoxby, 2000; Mosteller, 1995; Mueller, 2013; Nye et al., 2000; Shapson et al., 1980; Watson et al., 2013), it is more likely that schools will obtain benefits from attempting to keep class sizes from getting too large. Whereas some schools or school districts might be fortunate in evading the relationship between higher class sizes and lower academic performance, this case may be an exception rather than a rule (Cho et al., 2012; Glass, 1982; Goldstein & Blatchford, 1998; Hanushek, 1999; Hattie, 2005; Hoxby, 2000; Mosteller, 1995; Mueller, 2013; Nye et al., 2000; Shapson et al., 1980; Watson et al., 2013). The recommendation for the Mohawk Area School District is for class sizes not to exceed the upper range of 27 students in a class that was analyzed in this study. As class size increases the achievement is affected in a minor way but if class sizes were to increase substantially the outcome may negatively affect achievement.

Another recommendation is for the classroom teacher to utilize the regression model that evaluates actual i-Ready scores to determine whether the class is achieving the standard set forth by Mohawk classrooms with similar compositions. The classroom teacher would benefit from the knowledge of what the average classroom i-Ready score should be and then take proactive steps to push the students to exceed this mark to make achievement growth as a class.

Summary

School district officials have had difficulty making decisions about the optimal class size and its subsequent effect on student achievement to determine the most favorable class size environment for educators and students. The literature indicates that smaller class sizes are indicative of an increase in student achievement. The purpose of this study was to determine the effect of class size in the Mohawk Area School District on student academic achievement in reading and math across grade levels K through eight. The independent meta-analytic study determined that there was no meaningful effect of class size on student achievement performance in reading and math in the Mohawk Area School District.

Appendix A

Institutional Review Board (IRB)

Protocol #: 2018-002-88-A

Protocol Title: Examining the Effects of Class Size in the Mohawk Area School District: An independent Participant Data Meta-Analytic Investigation

The Institutional Review Board (IRB) of Slippery Rock University received the requested modifications to the above-referenced protocol.

The IRB has reviewed the modifications and approved the protocol under the EXEMPT category of review.

You may begin your project as of September 7, 2017.

Appendix B

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
12.000	1	515.170	9.042	81.756	497.448	532.892	56.976	0.000	0.000	0	1.000	0.000	0.000	0.000	0.000	0.000
14.000	2	573.496	6.342	40.226	561.065	585.927	90.423	0.000	22.536	1	0.000	95.561	1815.410	2686.645	7218060.61	42.608
15.000	5	540.611	3.960	15.678	532.851	548.372	136.535	0.000	76.355	4	0.000	94.761	1542.439	1248.215	1558041.46	39.274
16.000	7	554.984	4.200	17.640	546.752	563.216	132.138	0.000	294.049	6	0.000	97.960	6018.531	3676.617	13517512.8	77.579
17.000	5	539.677	3.804	14.474	532.221	547.134	141.853	0.000	197.721	4	0.000	97.977	3577.516	2669.449	7125957.17	59.812
18.000	8	507.854	2.758	7.604	502.449	513.259	184.165	0.000	811.816	7	0.000	99.138	7358.885	4458.449	19877766.9	85.784
19.000	10	484.779	2.263	5.119	480.345	489.214	214.256	0.000	1531.985	9	0.000	99.413	8997.921	4858.576	23605757.9	94.857
20.000	17	493.418	1.835	3.367	489.821	497.014	268.885	0.000	2015.780	16	0.000	99.206	7294.608	2928.660	8577051.61	85.408
21.000	7	579.662	3.399	11.552	573.000	586.323	170.550	0.000	128.391	6	0.000	95.327	1661.029	1022.406	1045314.94	40.756
22.000	10	582.600	2.529	6.398	577.643	587.558	230.325	0.000	162.685	9	0.000	94.468	1103.431	570.380	325333.606	33.218
23.000	16	559.742	1.729	2.990	556.352	563.131	323.701	0.000	1465.270	15	0.000	98.976	4713.705	1940.737	3766461.65	68.656
24.000	2	558.245	3.023	9.136	552.321	564.169	184.696	0.000	3.713	1	0.054	73.065	72.425	140.181	19650.810	8.510
25.000	4	625.162	2.814	7.921	619.646	630.678	222.134	0.000	39.612	3	0.000	92.426	409.308	383.294	146914.288	20.231
26.000	2	609.889	4.253	18.089	601.554	618.224	143.413	0.000	4.265	1	0.039	76.552	122.367	226.058	51102.295	11.062
27.000	1	577.560	4.913	24.134	567.931	587.189	117.565	0.000	0.000	0	1.000	0.000	0.000	0.000	0.000	0.000
Total within									6754.167	82	0.000					
Total between									3223.684	14	0.000					
Overall	97	544.211	0.754	0.568	542.733	545.689	722.017	0.000	9977.851	96	0.000	99.038	5699.447	993.706	987451.991	75.495

Appendix C

Groups	Effect size and 95% confidence interval						Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
0.000	12	-1.289	0.000	0.000	-1.290	-1.289	-6939.351	0.000	45.566	11	0.000	75.859
1.000	11	-0.561	0.031	0.001	-0.622	-0.500	-18.039	0.000	11.286	10	0.336	11.392
2.000	11	0.360	0.011	0.000	0.338	0.383	31.357	0.000	69.526	10	0.000	85.617
3.000	11	0.354	0.020	0.000	0.316	0.393	18.064	0.000	251.401	10	0.000	96.022
4.000	11	1.002	0.004	0.000	0.994	1.009	272.578	0.000	488.021	10	0.000	97.951
5.000	10	1.142	0.007	0.000	1.129	1.155	168.070	0.000	59.428	9	0.000	84.856
6.000	9	0.594	0.002	0.000	0.590	0.598	292.029	0.000	3944.695	8	0.000	99.797
7.000	11	1.562	0.046	0.002	1.470	1.653	33.602	0.000	8.058	10	0.623	0.000
8.000	11	1.562	0.048	0.002	1.467	1.656	32.469	0.000	12.311	10	0.265	18.774
Total within									4890.292	88	0.000	
Total between									1392537.93	8	0.000	
Overall	97	-1.266	0.000	0.000	-1.266	-1.265	-6852.277	0.000	1397428.22	96	0.000	99.993

Appendix D

Groups	Effect size and 95% confidence interval						Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
1.000	61	0.622	0.002	0.000	0.618	0.626	326.402	0.000	17054.584	60	0.000	99.648
2.000	25	0.952	0.004	0.000	0.944	0.959	262.085	0.000	9067.887	24	0.000	99.735
3.000	7	-1.289	0.000	0.000	-1.290	-1.289	-6938.964	0.000	554.811	6	0.000	98.919
4.000	4	0.075	0.059	0.004	-0.041	0.192	1.269	0.205	13.406	3	0.004	77.622
Total within									26690.688	93	0.000	
Total between									1370737.53	3	0.000	
Overall	97	-1.266	0.000	0.000	-1.266	-1.265	-6852.277	0.000	1397428.22	96	0.000	99.993
Mixed effects analysis												

Appendix E

Groups	Effect size and 95% confidence interval						Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
1.000	1	1.810	0.132	0.017	1.551	2.069	13.713	0.000	0.000	0	1.000	0.000
2.000	5	1.117	0.070	0.005	0.980	1.255	15.904	0.000	40.284	4	0.000	90.070
3.000	14	1.215	0.013	0.000	1.189	1.241	92.748	0.000	1835.992	13	0.000	99.292
4.000	25	0.199	0.019	0.000	0.162	0.236	10.502	0.000	4153.782	24	0.000	99.422
5.000	28	-1.288	0.000	0.000	-1.288	-1.287	-6932.080	0.000	124442.744	27	0.000	99.978
6.000	15	0.947	0.004	0.000	0.940	0.954	267.704	0.000	4234.111	14	0.000	99.669
7.000	9	0.569	0.002	0.000	0.565	0.573	279.086	0.000	3982.444	8	0.000	99.799
Total within									138689.357	90	0.000	
Total between									1258738.87	6	0.000	
Overall	97	-1.266	0.000	0.000	-1.266	-1.265	-6852.277	0.000	1397428.22	96	0.000	99.993

Appendix F

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
12.000	1	462.750	5.199	27.030	452.560	472.940	89.002	0.000	0.000	0	1.000	0.000	0.000	0.000	0.000	0.000
13.000	2	473.057	3.757	14.118	465.690	480.422	125.900	0.000	44.802	1	0.000	97.768	1256.062	1816.894	3301100.71	35.441
14.000	3	491.760	3.172	10.060	485.543	497.976	155.043	0.000	100.639	2	0.000	98.013	1720.318	1944.798	3782237.33	41.507
16.000	6	448.864	2.430	5.903	444.102	453.626	184.744	0.000	122.871	5	0.000	95.931	937.309	695.945	489922.904	30.616
17.000	3	447.415	3.383	11.443	440.785	454.045	132.261	0.000	167.305	2	0.000	98.805	2915.161	2978.912	8873919.32	53.992
18.000	11	473.514	1.485	2.206	470.603	476.425	318.830	0.000	1089.406	10	0.000	99.082	2652.866	1258.209	1583050.75	51.506
19.000	8	442.207	1.774	3.148	438.730	445.685	249.220	0.000	866.566	7	0.000	99.192	3211.565	1912.801	3658806.85	56.671
20.000	11	427.317	1.418	2.010	424.538	430.096	301.393	0.000	1129.908	10	0.000	99.115	2542.180	1260.388	1588578.80	50.420
21.000	8	501.040	1.656	2.743	497.794	504.286	302.526	0.000	583.952	7	0.000	98.800	1863.604	1093.287	1195476.74	43.193
22.000	8	471.348	1.596	2.546	468.221	474.476	295.422	0.000	178.973	7	0.000	96.089	502.104	282.237	79657.959	22.408
23.000	16	472.476	1.128	1.273	470.264	474.687	418.802	0.000	2825.057	15	0.000	99.469	3942.495	1716.286	2945601.79	62.789
24.000	10	515.141	1.295	1.676	512.604	517.679	397.872	0.000	270.442	9	0.000	96.672	515.324	296.913	88157.294	22.701
25.000	3	502.044	2.544	6.472	497.058	507.031	197.346	0.000	22.318	2	0.000	91.039	198.199	218.143	47586.210	14.077
26.000	2	537.198	2.428	5.896	532.439	541.957	221.228	0.000	104.356	1	0.000	99.042	1814.905	2591.611	6716440.21	42.603
27.000	1	513.810	3.072	9.439	507.788	519.832	167.241	0.000	0.000	0	1.000	0.000	0.000	0.000	0.000	0.000
Total within									7507.155	78	0.000					
Total between									3820.952	14	0.000					
Overall	93	476.926	0.482	0.232	475.981	477.871	989.220	0.000	11328.107	92	0.000	99.188	2650.734	458.365	210098.538	51.485

Appendix G

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
0.000	12	-1.705	0.023	0.001	-1.751	-1.660	-73.787	0.000	4.541	11	0.951	0.000
1.000	11	-1.213	0.002	0.000	-1.218	-1.209	-539.010	0.000	18.525	10	0.047	46.020
2.000	11	-0.832	0.020	0.000	-0.872	-0.792	-40.661	0.000	14.481	10	0.152	30.944
3.000	11	-0.559	0.001	0.000	-0.561	-0.556	-489.201	0.000	39.243	10	0.000	74.518
4.000	8	-0.399	0.018	0.000	-0.434	-0.363	-22.113	0.000	8.399	7	0.299	16.659
5.000	10	-0.215	0.035	0.001	-0.283	-0.147	-6.217	0.000	13.508	9	0.141	33.374
6.000	9	-0.361	0.005	0.000	-0.370	-0.351	-74.760	0.000	85.906	8	0.000	90.687
7.000	10	0.099	0.010	0.000	0.080	0.118	10.091	0.000	50.783	9	0.000	82.277
8.000	11	0.106	0.048	0.002	0.012	0.201	2.199	0.028	10.817	10	0.372	7.553
Total within									246.204	84	0.000	
Total between									80762.600	8	0.000	
Overall	93	-0.671	0.001	0.000	-0.673	-0.669	-680.006	0.000	81008.804	92	0.000	99.886

Appendix H

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
1.000	57	-1.142	0.002	0.000	-1.146	-1.138	-525.796	0.000	18794.928	56	0.000	99.702
2.000	21	-0.559	0.001	0.000	-0.561	-0.556	-490.239	0.000	922.556	20	0.000	97.832
3.000	9	-1.046	0.037	0.001	-1.117	-0.974	-28.627	0.000	366.260	8	0.000	97.816
4.000	6	-0.364	0.005	0.000	-0.374	-0.355	-75.560	0.000	8.962	5	0.111	44.209
Total within									20092.705	89	0.000	
Total between									60916.099	3	0.000	
Overall	93	-0.671	0.001	0.000	-0.673	-0.669	-680.006	0.000	81008.804	92	0.000	99.886

Appendix I

Groups	Effect size and 95% confidence interval						Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
2.000	4	-1.233	0.008	0.000	-1.249	-1.218	-153.180	0.000	94.484	3	0.000	96.825
3.000	14	0.051	0.042	0.002	-0.032	0.133	1.205	0.228	33.365	13	0.002	61.037
4.000	22	-0.550	0.001	0.000	-0.552	-0.548	-485.207	0.000	4665.160	21	0.000	99.550
5.000	29	-0.740	0.014	0.000	-0.767	-0.714	-54.329	0.000	2012.479	28	0.000	98.609
6.000	16	-1.049	0.002	0.000	-1.053	-1.044	-497.471	0.000	25194.130	15	0.000	99.940
7.000	8	-0.991	0.023	0.001	-1.036	-0.947	-43.614	0.000	195.434	7	0.000	96.418
Total within									32195.051	87	0.000	
Total between									48813.753	5	0.000	
Overall	93	-0.671	0.001	0.000	-0.673	-0.669	-680.006	0.000	81008.804	92	0.000	99.886

Appendix J

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
1.000	93	-0.671	0.001	0.000	-0.673	-0.669	-680.006	0.000	81008.804	92	0.000	99.886
2.000	97	-1.266	0.000	0.000	-1.266	-1.265	-6852.277	0.000	1397428.22	96	0.000	99.993
Total within									1478437.03	188	0.000	
Total between									350108.486	1	0.000	
Overall	190	-1.246	0.000	0.000	-1.246	-1.245	-6860.467	0.000	1828545.51	189	0.000	99.990
Mixed effects analysis												

Appendix K

Main results for Model 1, Random effects (MM), Z-Distribution, Mean

Covariate	Coefficient	Standard Error	95% Lower	95% Upper	Z-value	2-sided P-value
Intercept	-1.9188	0.4828	-2.8650	-0.9725	-3.97	0.0001
Class size	0.0519	0.0234	0.0060	0.0979	2.21	0.0268
Grade	0.3541	0.0277	0.2997	0.4085	12.76	0.0000

Statistics for Model 1

Test of the model: Simultaneous test that all coefficients (excluding intercept) are zero
Q = 172.64, df = 2, p = 0.0000
Goodness of fit: Test that unexplained variance is zero
Tau^2 = 0.4500, Tau = 0.6708, I^2 = 99.84%, Q = 58670.68, df = 94, p = 0.0000

Comparison of Model 1 with the null model

Total between-study variance (intercept only)
Tau^2 = 2.0033, Tau = 1.4154, I^2 = 99.99%, Q = 1397428.23, df = 96, p = 0.0000
Proportion of total between-study variance explained by Model 1
R^2 analog = 0.78

Number of studies in the analysis 97

Appendix L

Main results for Model 1, Random effects (MM), Z-Distribution, Mean

Covariate	Coefficient	Standard Error	95% Lower	95% Upper	Z-value	2-sided P-value
Intercept	-2.0467	0.1757	-2.3911	-1.7023	-11.65	0.0000
Class size	0.0307	0.0086	0.0138	0.0476	3.57	0.0004
Grade	0.2198	0.0097	0.2007	0.2389	22.58	0.0000

Statistics for Model 1

Test of the model: Simultaneous test that all coefficients (excluding intercept) are zero
$Q = 577.61$, $df = 2$, $p = 0.0000$
Goodness of fit: Test that unexplained variance is zero
$Tau^2 = 0.0347$, $Tau = 0.1863$, $I^2 = 96.24\%$, $Q = 2391.62$, $df = 90$, $p = 0.0000$

Comparison of Model 1 with the null model

Total between-study variance (intercept only)
$Tau^2 = 0.1927$, $Tau = 0.4390$, $I^2 = 99.89\%$, $Q = 81008.80$, $df = 92$, $p = 0.0000$
Proportion of total between-study variance explained by Model 1
R^2 analog = 0.82

Number of studies in the analysis 93

References

Abrantes, J. L., Seabra, C., & Lages, L. F. (2007). Pedagogical affect, student interest, and learning performance. *Journal of Business Research, 60*, 960-964.

Adkins-Coleman, T. A. (2010). "I'm not afraid to come into your world": Case studies of teachers facilitating engagement in urban high school English classrooms. *The Journal of Negro Education, 79*(1), 41-53.

Ahmed, W., Minnaert, A., Kuyper, H., & van der Werf, G. (2012). Reciprocal relationships between math self-concept and math anxiety. *Learning and Individual Differences, 22*(3), 385-389.

Altman, D. G. (1991). *Practical statistics for medical research*. London, U.K.: Chapman and Hall.

Anderson, E. (2012). Reflections on the "Black-White achievement gap". *JOURNAL OF SCHOOL PSYCHOLOGY, 50*(5), 593-597. doi:10.1016/j.jsp.2012.08.007

Bagnardi, V., Blangiardo, M., La Vecchia, C., & Corrao, G. (2001). A meta-analysis of alcohol drinking and cancer risk. *British journal of cancer, 85*(11), 1700-1705.

Balnaves, M., & Caputi, P. (2012). *Introduction to quantitative research methods: An investigative approach*. Thousand Oaks, CA: Sage.

Blondal, K. S., & Adalbjarnardottir, S. (2012). Student disengagement in relation to expected and unexpected educational pathways. *Scandinavian Journal of Educational Research, 56*(1), 85-100.

Bobbitt-Zeher, D. (2004). Black American students in an affluent suburb: A study of academic disengagement (Vol. 33, pp. 414-416). Washington, D.C.: American Sociological Association.

Bondy, E., Ross, D. D., Hambacher, E., & Acosta, M. (2013). Becoming warm demanders: Perspectives and practices of first year teachers. *Urban Education, 48*(3), 420-450. doi:10.1177/0042085912456846

Bonner, E. P. (2014). Investigating practices of highly successful mathematics teachers of traditionally underserved students. *Educational Studies in Mathematics, 86*(3), 377-399. doi:10.1007/s10649-014-9533-7

Bonner, E. P., & Adams, T. L. (2012). Culturally responsive teaching in the context of mathematics: a grounded theory case study. *Journal of Mathematics Teacher Education, 15*(1), 25-38. doi:10.1007/s10857-011-9198-4

Bonta, J., Blais, J., & Wilson, H. A. (2014). A theoretically informed meta-analysis of the risk for general and violent recidivism for mentally disordered offenders. *Aggression and Violent Behavior, 19*(3), 278-287. doi:10.1016/j.avb.2014.04.014

Borenstein, M., Hedges, L. V., Higgings, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, U.K.: Wiley.

Borg, J. R., Mary, O. B., & Harriet, A. S. (2012). Closing the achievement gap between high-poverty schools and low-poverty schools. *Research in Business and Economics, 5*, 1-10.

Bouta, H., Retalis, S., & Paraskeva, F. (2012). Utilising a collaborative macro-script to enhance student engagement: A mixed method study in a 3D virtual environment. *Computers & Education, 58*(1), 501-517.

Burchinal, M., McCartney, K., Steinberg, L., Crosnoe, R., Friedman, S. L., McLoyd, V., & Pianta, R. (2011). Examining the black–white achievement gap among low-income children using the NICHD Study of Early Child Care and Youth Development. *Child development, 82*(5), 1404-1420.

Burkam, D., & Lee, V. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Economic Policy Institute.

Burton, E., Cavalheri, V., Adams, R., Browne, C. O., Bovery-Spencer, P., Fenton, A. M., . . . Hill, K. D. (2015). Effectiveness of exercise programs to reduce falls in older people with dementia living in the community: a systematic review and meta-analysis. *Clinical interventions in aging, 10*, 421-434.

Campbell, D. T. (1988). *Methodology and epistemology for social sciences: Selected papers*. Chicago, IL: University of Chicago Press.

Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review, 31*(3), 77-95.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning a systematic review and meta-analysis. *Review of educational research, 86*(1), 79-122.

Cochrane, A. L. (1972). *Effectiveness and efficiency*. Nuffield, U.K.: The Nuffield Provincial Hospitals Trust.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge.

Cole, M. G., & Dendukuri, N. (2014). Risk factors for depression among elderly community subjects: a systematic review and meta-analysis. *American Journal of Psychiatry, 160*(6), 1147-1156.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Creswell, J. W. (2015). *Research methods*. Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Csörgo, M., & Révész, P. (2014). *Strong approximations in probability and statistics*. New York, NY: Academic Press.

Cuijpers, P., Weitz, E., Karyotaki, E., Garber, J., & Andersson, G. (2015). The effects of psychological treatment of maternal depression on children and parental functioning: a meta-analysis. *European child & adolescent psychiatry, 24*(2), 237-245.

Cummins, J. (2009). Transformative multiliteracies pedagogy: School-based strategies for closing the achievement gap. *Multiple Voices for Ethnically Diverse Exceptional Learners, 11*(2), 38-56.

Davies, M. B., & Hughes, N. (2014). *Doing a successful research project: Using qualitative or quantitative methods*. New York, NY: Palgrave Macmillan.

Davis, C. (2012). Achievement gap woes. *Diverse Issues in Higher Education, 29*(17), 5-5.

Davis, H. A., Gabelman, M. M., & Wingfield, R. D. (2011). " She let us be smart:" Low-income African-American first-grade students' understandings of teacher closeness and influence. *The Journal of Classroom Interaction*, 4-16.

Dickson, R., & Cherry, M. G. (2014). Defining my review question and identifying inclusion criteria. In R. Dickson, M. G. Cherry, & A. Boland (Eds.), *Doing a systematic review: A student's guide* (pp. 17-34). London, U.K.: Sage.

Dixon, N. M. (2000). *Common knowledge: How companies thrive by sharing what they know*. Cambridge, MA: Harvard Business Press.

Dodson, A. P. (2012). Coming to the defense of black students. *Diverse Issues in Higher Education, 29*(6), 21-21.

Durham, J. (2012). Examining the achievement gap with a psychodynamic lens: Implications for practice. *Journal of Infant, Child, and Adolescent Psychotherapy, 11*(3), 217. doi:10.1080/15289168.2012.700795

Fan, W., & Williams, C. M. (2010). The effects of parental involvement on students' academic self-efficacy, engagement and intrinsic motivation. *Educational Psychology, 30*(1), 53-74.

Ford, A. C., & Sassi, K. (2014). Authority in cross-racial teaching and learning (Re)considering the transferability of warm demander approaches. *Urban Education, 49*(1), 39-74. doi:10.1177/0042085912464790

Franz, M. J., Boucher, J. L., Rutten-Ramos, S., & VanWormer, J. J. (2015). Lifestyle weight-loss intervention outcomes in overweight and obese adults with type 2 diabetes: a systematic review and meta-analysis of randomized clinical trials. *Journal of the Academy of Nutrition and Dietetics, 115*(9), 1447-1463.

Gans, J., Replogle, E., & Tichenor, D. (2008). *Debates on U.S. Immigration*. Thousand Oaks: Sage.

Ghazvini, S. D. (2011). Relationships between academic self-concept and academic performance in high school students. *Procedia-Social and Behavioral Sciences, 15*, 1034-1039.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Glass, G. V. (1982). *School class size: Research and policy*. Beverly Hills, CA: Sage.

Go, S., & Lindert, P. (2010). The uneven rise of American public schools to 1850. *The Journal of Economic History, 70*(01), 1-26.

Goldstein, H., & Blatchford, P. (1998). Class size and educational achievement: a review of methodology with particular reference to study design. *British Educational Research Journal, 24*(3), 255-268.

Gordon, B. M. (2012). “Give a brotha a break!”: The experiences and dilemmas of middle-class African American male students in white suburban schools. *TEACHERS COLLEGE RECORD, 114*(5), 1-26.

Grantham, T. C., & Biddle, W. H. (2014). From bystander to upstander: Teacher for gifted black students accused of acting white. *Gifted Child Today, 37*(3), 178-187.

Green, J., Liem, G. A. D., Martin, A. J., Colmar, S., Marsh, H. W., & McInerney, D. (2012). Academic motivation, self-concept, engagement, and performance in high school: Key processes from a longitudinal perspective. *Journal of adolescence, 35*(5), 1111-1122.

Griner, A. C., & Stewart, M. L. (2013). Addressing the achievement gap and disproportionality through the use of culturally responsive teaching practices. *Urban Education, 48*(4), 585-621. doi:10.1177/0042085912456847

Hall Mark, D. L. (2013). Academic achievement gap or gap of opportunities? *Urban Education, 48*(2), 335-343.

Hampden-Thompson, G., Guzman, L., & Lippman, L. (2013). A cross-national analysis of parental involvement and student literacy. *International Journal of Comparative Sociology, 54*(3), 246-266. doi:10.1177/0020715213501183

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21*(2), 143-163.

Hartney, M. T., & Flavin, P. (2014). The political foundations of the Black–White education achievement gap. *American Politics Research, 42*(1), 3-33. doi:10.1177/1532673X13482967

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research, 43*(6), 387-425.

Hayes, D. (2012). Parental involvement and achievement outcomes in African American adolescents. *Journal of Comparative Family Studies, 43*(4), 567-582.

Henderikus, S. (2010). Theory. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1498-1502). Thousand Oaks, CA: Sage.

Hing, B. O. (1994). *Making and remaking Asian America through immigration policy: 1850-1990*. Palo Alto: Stanford University Press.

Houchen, D. (2013). "Stakes is high": Culturally relevant practitioner inquiry with African American students struggling to pass secondary reading exit exams. *Urban Education, 48*(1), 92-115. doi:10.1177/0042085912456845

Howes, C., & Ritchie, S. (2002). *A matter of trust: connecting teachers and learners in the early childhood classroom*. New York, NY: Teachers College.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics, 115*(4), 1239-1285.

Izzo, R. L., & Ross, R. R. (1990). Meta-analysis of rehabilitation programs for juvenile delinquents: A brief report. *Criminal Justice and Behavior, 17*(1), 134-142.

Jackson, S. (2015). *Research methods and statistics: A critical thinking approach.* New York, NY: Cengage Learning.

Jeynes, W. (2012). A meta-analysis of the efficacy of different types of parental involvement programs for urban students. *Urban Education, 47*(4), 706-742. doi:10.1177/0042085912445643

Keppel, G., Saufley, W. H., & Tokunaga, H. (1992). *Introduction to design and analysis.* New York, NY: Macmillan.

Klassen, R. M., & Chiu, M. M. (2011). The occupational commitment and intention to quit of practicing and pre-service teachers: Influence of self-efficacy, job stress, and teaching context. *Contemporary Educational Psychology, 36*(2), 114-129. doi:10.1016/j.cedpsych.2011.01.002

Kremelberg, D. (2010). *Practical statistics: A quick and easy guide.* Thousand Oaks, CA: Sage.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review, 73*(1), 31-43.

Leary, M. R. (2011). *Introduction to behavioral research methods.* New York, NY: Pearson.

Lee, C. S., Hayes, K. N., Seitz, J., DiStefano, R., & O'Connor, D. (2016). Understanding motivational structures that differentially predict engagement and achievement in middle school science. *International Journal of Science Education, 38*(2), 1-24.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage Publications.

Luebke, T., & Brunkwall, J. (2015). Risk-adjusted meta-analysis of 30-day mortality of endovascular versus open repair for ruptured abdominal aortic aneurysms. *Annals of vascular surgery, 29*(4), 845-863.

Madigan, S., Atkinson, L., Laurin, K., & Benoit, D. (2013). Attachment and internalizing behavior in early childhood: A meta-analysis. *Developmental psychology, 49*(4), 672-689.

Matta, B. N., Guzman, J. M., Stockly, S. K., & Widner, B. (2015). Class Size Effects on Student Performance in a Hispanic-Serving Institution. *The Review of Black Political Economy, 42*(4), 443-457.

McBurney, D., & White, T. (2011). *Research methods*. New York, NY: Cengage.

McGraw, K. O., & Wong, S. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361-365.

McNabb, D. E. (2010). *Research methods for political science*. Thousand Oaks, CA: Sage.

Mo, Y., Singh, K., & Chang, M. (2013). Opportunity to learn and student engagement: a HLM study on eighth grade science achievement. *Educational Research for Policy and Practice, 12*(1), 3-19.

Moore, D. S., & McCabe, G. (2009). *Introduction to the practice of statistics*. New York, NY: W.H. Freeman.

Morris, S. B. (2008). Estimating effect sizes from the pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364-386.

Morrison, C., Marschark, M., Sarchet, T., Convertino, C. M., Borgna, G., & Dirmyer, R. (2013). Deaf students' metacognitive awareness during language comprehension. *European Journal of Special Needs Education, 28*(1), 78-90.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of children*, 113-127.

Moyer-Packenham, P., Salkind, G., Bolyard, J., & Suh, J. M. (2013). Effective choices and practices: Knowledgeable and experienced teachers' uses of manipulatives to teach mathematics. *Online Journal of Education Research, 2*(2), 18-33.

Mueller, S. (2013). Teacher experience and the class size effect—Experimental evidence. *Journal of Public Economics, 98*, 44-52.

Natrella, M. G. (2013). *Experimental statistics.* New York, NY: Courier Corporation.

Ngai, M. M. (1999). The architecture of race in American immigration law: A reexamination of the Immigration Act of 1924. *The Journal of American History, 86*(1), 67-92.

Nunnery, J. A., Chappell, S., & Arnold, P. (2013). A meta-analysis of a cooperative learning models effects on student achievement in mathematics. *Cypriot Journal of Educational Sciences, 8*(1), 34-48.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37*(1), 123-151.

Phillippo, K. (2012). "You're trying to know me": Students from nondominant groups respond to teacher personalism. *The Urban Review, 44*(4), 441-467. doi:10.1007/s11256-011-0195-9

Plunkett, S. W., Behnke, A. O., Sands, T., & Choi, B. Y. (2009). Adolescents' reports of parental engagement and academic achievement in immigrant families. *Journal of Youth and Adolescence, 38*(2), 257-268.

Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., & Madensen, T. D. (2006). The empirical status of deterrence theory: A meta-analysis. *Taking stock: The status of criminological theory, 15*, 367-396.

Price, D. V., & Tovar, E. (2014). Student engagement and institutional graduation rates: Identifying high-impact educational practices for community colleges. *Community College Journal of Research and Practice, 38*(9), 766-782.

Rios, V. (2012). Reframing the achievement gap. *Contexts, 11*(4), 8-10. doi:10.1177/1536504212466324

Roberts, T. G., & Dyer, J. E. (2004). Characteristics of effective agriculture teachers. *Journal of Agricultural Education, 45*, 82-95.

Rodriguez, R. J., & Elbaum, B. (2014). The Role of Student–Teacher Ratio in Parents' Perceptions of Schools' Engagement Efforts. *The Journal of Educational Research, 107*(1), 69-80.

Rojas-LeBouef, A., & Slate, J. R. (2012). The achievement gap between white and non-white students. *International Journal of Educational Leadership Preparation, 7*(1).

Roodman, D. (2007). The anarchy of numbers: aid, development, and cross-country empirics. *The World Bank Economic Review, 21*(2), 255-277.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York, NY: Sage.

Ross, D. D., Bondy, E., Gallingane, C., & Hambacher, E. (2008). Promoting academic engagement through insistence: Being a warm demander. *Childhood Education [H.W. Wilson - EDUC], 84*(3), 142.

Salkovsky, M., & Romi, S. (2015). Teachers' coping styles and factors inhibiting teachers' preferred classroom management practice. *Teaching and teacher education, 48*, 56-65.

Salvia, J., Ysseldyke, J., & Bolt, S. (2012). *Assessment: In special and inclusive education*. New York, NY: Cengage Learning.

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., . . . Woods, J. (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. *Computers & Education, 72*, 271-291.

Senler, B., & Sungur, S. (2009). Parental influences on students' self-concept, task value beliefs, and achievement in science. *The Spanish journal of psychology, 12*(01), 106-117.

Shapson, S. M., Wright, E. N., Eason, G., & Fitzgerald, J. (1980). An experimental study of the effects of class size. *American Educational Research Journal, 17*(2), 141-152.

Sun, J. C.-Y. (2014). Influence of polling technologies on student engagement: An analysis of student motivation, academic performance, and brainwave data. *Computers & Education, 72*, 80-89.

Tabachnick, B. G., and Fidell, L. S. (2013). Using Multivariate Statistics , 6th ed. Boston : Pearson.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research articles: A simplified methodology. Retrieved from http://www.bwgriffin.com/gsu/courses/edur9131/content/Effect_Sizes_pdf5.pdf

Theron, L. C. (2013). Black students' recollections of pathways to resilience: Lessons for school psychologists. *School Psychology International, 34*(5), 527-539. doi:10.1177/0143034312472762

Tobias, D. K., Chen, M., Manson, J. E., Ludwig, D. S., Willett, W., & Hu, F. B. (2015). Effect of low-fat diet interventions versus other diet interventions on long-term weight change in adults: a systematic review and meta-analysis. *The Lancet Diabetes & Endocrinology, 3*(12), 968-979.

Trochim, W., Donnelly, J., & Arora, K. (2015). *Research methods: The essential knowledge base*. Boston, MA: Nelson Education.

Usher, E., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology, 34*(1), 89-101. doi:10.1016/j.cedpsych.2008.09.002

Verhulst, B., Neale, M., & Kendler, K. (2015). The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. *Psychological medicine, 45*(05), 1061-1072.

Vogt, W. P., & Johnson, R. B. (2011). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. Thousand Oaks, CA: Sage.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174-1204.

Watson, K., Handal, B., Maher, M., & McGinty, E. (2013). Globalising the class size debate: myths and realities. *Journal of International and Comparative Education, 2*(2), 72-81.

Williams, T. T., & Sánchez, B. (2012). Parental involvement (and uninvolvement) at an inner-city high school. *Urban Education, 47*(3), 625-652. doi:10.1177/0042085912437794

Xu, J. Z., Coats, L. T., & Davidson, M. L. (2012). Making science homework work: The perspectives of exemplary African American science teachers. *TEACHERS COLLEGE RECORD, 114*(7), 1.

Yin, R. K. (2009). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

Zhang, H., Zhang, Y., & Seiler, M. J. (2013). Using behavioral and experimental design methods to teach real estate: Understanding asset pricing bubbles. *Journal of Real Estate Practice and Education, 16*(2), 107-118.

Zikmund, W. G. (2003). *Business research methods*. Mason, OH: Southwestern.