A Spatial Regression Discontinuity Evaluation of Minnesota's Quality Compensation for Teachers Program

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

Christopher Thomas Moore

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

Advisers: Ernest C. Davenport Jr. and Frances P. Lawrenz

June 2015

Acknowledgments

I owe so much to my teachers, from kindergarten through graduate school. Thank you for selflessly dedicating yourselves to educating others. I am especially grateful to my advisers, Dr. Ernest Davenport and Dr. Frances Lawrenz--I look up to you and feel so fortunate that you believed in me. I am also grateful to my committee members and others who reviewed my dissertation research and provided valuable feedback along the way: Dr. Morris Kleiner, Dr. Andrew Zieffler, Dr. Michael Harwell, Dr. Aaron Sojourner, Dr. Elton Mykerezi, Dr. Kristine West, and Ann Mavis. I also wish to thank Dr. Susan Ambler and Dr. Jeff Bay who mentored me and sparked my interest in research as an undergraduate at Maryville College.

Dedication

I dedicate this dissertation to my wife and parents. To my wonderful wife, Amy Wick Moore, I love you and thank you for your encouragement and patience. I am looking forward to more free time and adventures together. To my parents, Linda and Steve Moore, thank you for teaching me the importance of hard work, education, and helping others. I love you and aspire to live by your example.

Abstract

The Quality Compensation for Teachers (Q Comp) program provides up to $260 per student to Minnesota schools that adopt reforms to teacher pay and professional development. The reforms include an alternative salary schedule and observations of classroom instruction. Q Comp participants received about $419 million between 2006-2013, but voluntary participation and variability in implementation have made it challenging to evaluate Q Comp's overall impact on student achievement and identify its most effective reforms. This study applies spatial regression discontinuity (RD) and other quasi-experimental methods to estimate the effect of Q Comp participation and identify exemplars. Participation is estimated to significantly increase math and reading achievement by 0.0541 and 0.0247 standard deviation, respectively, compared to geographically neighboring districts that did not participate. The estimates are robust and diverge from a nonequivalent dependent variable. School distance from the Q Comp border is not a significant RD assignment variable. Five participating districts (Farmington, North St. Paul-Maplewood, Osseo, Spring Lake Park, and St. Francis) exhibited achievement gains that were significantly larger than expected, making them good candidates for qualitatively investigating which Q Comp reforms are most effective.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The State of Minnesota enacted a program in 2005 to reform teacher pay and professional development. School districts and charter schools voluntarily participate in the program, named Quality Compensation for Teachers (Q Comp), in exchange for additional per-pupil funding. Four hundred nineteen million dollars have been allotted to 70 districts and $22 million to 61 charter schools through the school year ending 2013. Participating districts and charters are required to add a variable-pay component to the traditional steps and lanes salary schedule and to enhance teachers' professional development and career advancement options. Q Comp's reforms are relatively new and untested, and they incorporate politically opposing views on how to raise the quality of instruction. The Minnesota Department of Education (MDE) interprets and implements the statutes authorizing Q Comp.

By what criteria and methods should Q Comp be evaluated? Fitzpatrick, Sanders, and Worthen (2004) define program evaluation as "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit)" (p. 5). This evaluation study applies an *objectives-oriented* approach to statistically estimate Q Comp's impact on student achievement. Several factors warrant evaluating Q Comp in such a manner:

1. its scale of participation and expenditures
2. the relatively new and politically contentious educational policies it represents
3. voluntary participation at the district level
4. Q Comp's stated goal "to improve student learning" makes student achievement an important criterion for judging its effectiveness.

Other evaluation approaches, methods and outcomes are not feasible within this study but would be especially helpful for distilling and articulating Q Comp's most effective components for utilization by a variety of Q Comp stakeholders.

Estimating causal effects is an important aim in objectives-oriented evaluations, and randomized field experiments are considered the gold standard (Boruch, 1991). However, many programs and policies are implemented without randomly assigning participants to a treatment or control group. Regression discontinuity (RD) design is a quasi-experimental approach that is widely applicable for program evaluation. RD (by definition) and programs (out of convenience) assign participants to a treatment (e.g., tutoring) based on a sharp cutoff point (e.g., poverty criterion) along a continuous assignment variable (e.g., income; Imbens & Lemieux, 2007). In contrast to a randomized experiment in which the assignment variable is assumed to be independent of all other variables, the RD treatment assignment mechanism is *determined* by a continuous pre-test variable. Piecewise regression analysis (a.k.a., broken stick, segmented, or switching regression) is used to estimate local average treatment effects (LATEs) at the cutoff. RDs are favored by federal agencies and evaluation theorists because, when the selection process is completely known, LATE estimates are comparable to non-local estimates from randomized controlled trials (U.S. Department of Education, 2005; Cook, 1991; Cook, Shadish, & Wong, 2008).

Like Q Comp, many programs and policies are implemented in geographically defined jurisdictions, such as school districts or states, and without random assignment. How might evaluators estimate causal effects in the case of treatment assignment based on geographic borders? This study describes *spatial regression discontinuity* (SRD)

design and analysis and applies SRD to evaluate Q Comp. SRD is a variation of traditional, univariate RD that recognizes geographic borders as sharp cutoff points where geographically local average treatment effects (GLATEs) of programs and policies can be estimated (Holmes, 1998; Black, 1999). SRD shares some of the elements of RD that strengthen internal and external validity, but GLATEs estimated at preexisting geographic borders are subject to a number of validity threats that can be ameliorated through design and enhancements. Voluntary participation at the district level makes it difficult to conduct a defensible objectives-oriented evaluation of Q Comp. I attempt to address the validity threats by applying SRD theory and best practices. It should be stressed that although this dissertation describes SRD, integrates related theories, and demonstrates best practices, it does not present a new methodology as much as contribute to evaluation practice, especially as it relates to evaluating Q Comp and other educational policies implemented by a subset of school districts within a state. That is, a goal of this dissertation is to demonstrate best practices in a manner that enables other researchers to understand SRD, apply the methodology to other real-world or simulated situations, validate findings, and contribute to a better understanding of SRD.

Chapter 2: Literature review

This chapter presents an overview of Minnesota's Quality Compensation for Teachers (Q Comp) program and discusses approaches to evaluating its impact on student achievement. It introduces spatial regression discontinuity (SRD) and summarizes validity threats and methodological enhancements for ruling out validity threats. It concludes with a discussion of why Q Comp qualifies as a spatial regression discontinuity (SRD) natural experiment.

Minnesota's Quality Compensation for Teachers (Q Comp) program

In 2005, the State of Minnesota established a "restructured teacher compensation system ... to provide incentives for teachers to improve their knowledge and skills and for school districts to recruit and retain highly qualified teachers, and support teachers' roles in improving students' educational achievement" (*Laws of Minnesota*, 2005). The language of the law was later revised to emphasize two additional purposes: "to improve student learning [and] encourage highly qualified teachers to undertake challenging assignments" (*Minnesota Statutes*, 2008). The Minnesota Department of Education (MDE), charged with interpreting and implementing the law authorizing the program, named it Quality Compensation for Teachers, or Q Comp.

Q Comp is a voluntary program that incentivizes school districts and charter schools to participate by promising additional funding. Participating school districts automatically received $260 per student in additional state revenue in 2006, $190.06 in 2007-2009 and $169 in 2010-present. Q Comp legislation has permitted districts to levy additional revenue up to $260 (i.e., $69.94 in 2007-2009 and $91 in 2010-present). Interested districts and charter schools must design their own plan and have it approved

by MDE and teachers in their district or charter school. In the case of districts in which teachers are represented by a union, teachers formally vote on the question of Q Comp participation. In order to receive approval from MDE, a district or charter school's plan must include five components:

1. career advancement options
2. ongoing professional development during the school day
3. frequent observation of instruction by trained raters
4. performance pay for student achievement and instructional ratings
5. an alternative salary schedule.

The alternative salary schedule component of Q Comp requires districts and charter schools to reform the traditional teacher salary schedule in which years of experience and levels of education are the only determinants of a teacher's pay. The traditional schedule is also known as the "single salary schedule," "position-automatic schedule," and "steps and lanes schedule" (Office of the Legislative Auditor, 2009; Podgursky & Springer, 2007). It takes the form of a matrix with rows/steps representing the number of years of experience and columns/lanes representing graduate coursework or degrees obtained.

To comply with the performance pay component of Q Comp, districts and charter schools must provide performance pay to teachers for school-wide student achievement gains on a standardized test, a teacher's own classroom-level gains, ratings from their instructional observations, or other criteria, such as active participation in professional learning communities (PLCs). A district or charter school's plan specifies the degree to which each component determines performance pay. The Office of the Legislative

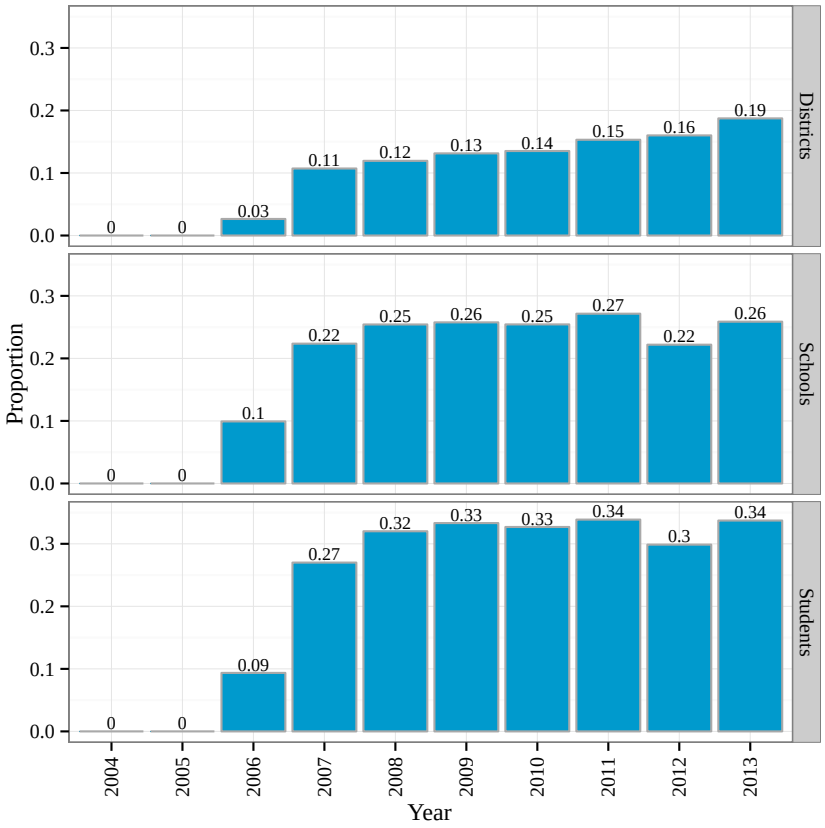
Auditor (2009) found that Q Comp districts and charter schools have typically tied about half of teacher performance pay to instructional ratings and divided the other half between school-level student gains, classroom-level gains, and other components. As of fall 2008, individual teacher performance pay awards had ranged from $68 to $2,500. Sojourner, Mykerezi, and West (in press) reviewed Q Comp applications and found that annual potential bonuses averaged $1,107 for classroom-observations, $850 for meeting student learning goals at the teacher- or grade-level (not typically measured by standardized tests), and $234 for meeting school- or district-wide learning goals measured by standardized tests.

Regular PLC meetings are a key component of Q Comp. PLCs are small teams of teachers that work together to improve their instructional skills. PLCs typically have a team leader who sets the agendas, runs the meetings, and serves as a resource for the other teacher members. The Office of the Legislative Auditor (2009) found that Q Comp teachers typically *advance* their careers by leading PLCs in their schools in exchange for a stipend of $150 to $3,500. And PLCs tend to be the primary source of ongoing professional development required by Q Comp. Q Comp teachers receive three observations a year by PLC leaders and/or other trained raters, including fellow teachers and principals.

Figure 2.1 shows Q Comp yearly participation rates for districts, schools and students in this study, and Appendix 1 shows Q Comp's geographic expansion over time. Note that for reasons described in the Data section below, participation rate denominators are Independent, Common and Special school districts; schools in those districts with 10 or more students in a cohort who took Minnesota's accountability tests for students with

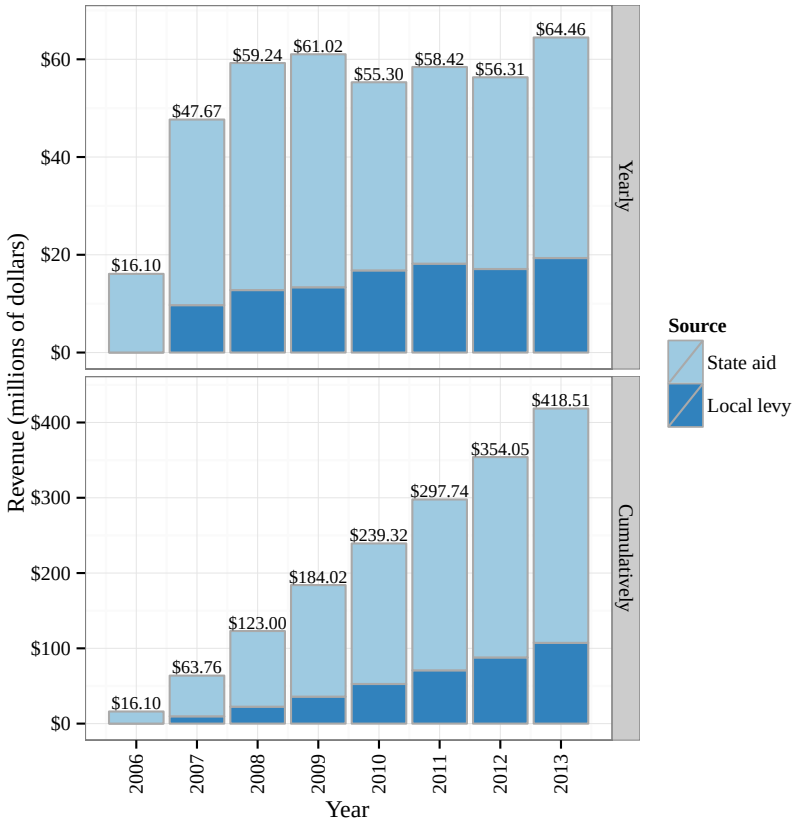
either no or mild cognitive disabilities; and those students tested in those schools. Only a handful of districts chose to participate in the program's first school year (fall 2005 - spring 2006; denoted 2006). District participation jumped the following year and has steadily increased, reaching 19 percent in 2013. Participation by schools and students has followed a similar trend, although their rates have not increased monotonically due to participation turnover among districts of different sizes. Larger districts have participated at a higher rate than smaller districts, which is evidenced by school and student participation rates that are higher than district rates. Larger districts were more likely to apply because they could justify the effort/fixed cost of applying in exchange for the per-student funding (Sojourner et al., in press). About 20 percent of Minnesota school districts (70 out of 345) and about 32 percent of public schools (557 out of 1,758) have participated in Q Comp at one time or another between fall 2005 and spring 2013 (Minnesota Department of Education, 2014; T. Yetter, personal communication, April 17, 2014).

Figure 2.1. Q Comp participation rates: Public school districts, schools and tested students



Participating school districts have received about $52 million a year for a total of $419 million in Q Comp revenue through the school year ending 2013 (see Figure 2.2). Of the $52 million, about $39 million has come from state aid each year for a total of $311 million. Districts have levied an additional $15 million a year since 2007 for a total of $107 million. (Sixty-one charter schools have received a total of $22 million in state aid and cannot levy locally.)

Figure 2.2. Q Comp revenue over time: Public school districts



Q Comp in context

Q Comp is part of a broader shift away from traditional teacher compensation. One reason for the shift is accumulating evidence that experience and education do not strongly influence student achievement (Hanushek, 2003). According to Podgursky and Springer (2007), the traditional salary schedule emerged in the 1920's in response to nepotism and gender and racial disparities in teacher pay, as well as indirectly from broader conflicts in industrial relations at that time. The private sector uses salary schedules to figure *base pay,* but unlike teachers, private sector workers can earn *variable pay* for reaching performance milestones. Recent experiments with teacher

compensation have linked variable pay to performance measured by *inputs* (i.e., knowledge and skills such as professional development) and *outputs* (i.e., merit such as student achievement gains). In 1999, Denver Public Schools began linking teacher pay to teacher evaluations, professional development, and student achievement. In the same year, the National Institute for Excellence in Teaching began administering the Teacher Advancement Program (TAP), which became a model for the Q Comp program (Office of the Legislative Auditor, 2009). Five districts participated in a pay-for-performance pilot starting in 2004, and Waseca schools and three Minneapolis schools piloted TAP starting in 2005 (Sojourner et al., in press). Texas rolled out a large performance pay program in 2006. Florida did so in 2007, and its legislature has considered tying half of all teachers' pay to student achievement measures (Alvarez, 2011). At the Federal level, $99 million dollars were appropriated in 2006 to the Teacher Incentive Fund, which provides competitive grants to localities to implement TAP and other performance pay programs. Eclipsing the 2006 appropriation, the American Recovery and Reinvestment Act of 2009 provided $4.35 billion in Race to the Top grants to implement teacher performance pay and other educational reforms (U.S. Department of Education, 2011).

Does performance pay work? Camerer and Hogarth (1999) and Prendergast (1999) reviewed decades of experimental and observational studies and conclude that, in general, incentives do work. Evidence supports theories that incentives help elicit greater effort of current employees and attract talented employees. *Incentives* generally refer to variable performance pay for individuals, but a broad interpretation of performance pay may include paying for group performance, profit sharing, higher-than-average pay levels (relative to competitors), deferred compensation, and promotion/career advancement.

Principal-agent theory helps explain why incentives can be effective. Principals (i.e., employers) and agents (i.e., employees) have different priorities, so principals attempt to optimize compensation to encourage behavior that benefits the firm without overpaying, and employees will seek to apply just enough effort to match pay. Emerging *human capital* theory recognizes that performance depends on cognitive ability (i.e., not just effort), that performance and effort are substitutes, and that incentives influence effort more than ability. Even though cognitive ability is insensitive to performance pay, having a performance pay and/or professional development system in place may raise human capital by attracting new employees with high potential and/or ability and retaining employees that have exhibited high ability.

Camerer and Hogarth (1999) and Prendergast (1999) qualify their endorsements of incentives when highly cognitively demanding jobs are held by intrinsically motivated employees who are risk averse and who must work as a team (e.g., teachers). They offer the following qualifications.

1. Performance is difficult to measure reliably, and operationalizing performance may unintentionally cause employees to focus on a small subset of tasks (e.g., teaching to the test).
2. Incentives can reduce productivity by reducing intrinsic motivation or by causing excessive effort so that it inhibits ability.
3. Incentives are risky to the degree that they account for total compensation, and perceptions of risk may overwhelm potential incentives to the point that highly able employees select out of the job.

4. When group performance is measured and rewarded, it may cause individuals to perform at a lower level compared to individual performance pay and may raise performance of individuals who respond to peer pressure, the net effect of which may be a reduction in performance *variability* without raising mean performance.

Camerer and Hogarth (1999) stress that redesigning tasks to match employees' cognitive abilities and focusing on employees' professional development may be more effective than paying for performance.

Designers of Q Comp embedded several elements of principal-agent and human capital economic theories (i.e., incentives and professional development) at different levels of implementation. In general, incentives directly impact effort but not ability. The same can be said for Q Comp. The State of Minnesota acts as a principal by offering incentives for districts to participate in Q Comp. Districts qualify as agents when it comes to applying for and complying with Q Comp requirements. Additional state aid revenue provided by Q Comp was cited by district leaders as a key incentive for submitting an application (i.e., to direct more pay to teachers without having to raise taxes on local residents), and many districts choose to impose the optional Q Comp levy permitted by the legislation in order to maximize per-student funding at $260 (Schwartz, 2012). As discussed extensively by Schwartz (2012), participating in Q Comp does not clearly signal the *ability* of a school district to implement authentic instructional reforms. Districts qualify as principals when it comes to variably paying teachers. That is, Q Comp incentives are invariant to *district* performance. They receive a fixed amount of revenue regardless of teacher and school performance but pay teachers for professional development effort and performance. The degree to which districts are able to reliably

measure and pay for performance remains unclear despite compliance with program requirements. Q Comp strongly emphasizes career advancement, PLCs and feedback about instructional practice, but of the many Q Comp professional development practices it remains unclear which ones and in which combination actually raise teachers' abilities and student achievement.

Schwartz (2012) uses the term "teacher improvement system" (TIS) to refer to Q Comp, the Teacher Advancement Program (TAP), Race to the Top, and similar district-level initiatives that attempt to transform and align incentives and professional development. Incentives and professional development are supposed to interact to accelerate student achievement more than either component could accomplish on its own, even though economic theory asserts that effort and human capital are substitutes. Q Comp funding and oversight encourages districts to adopt pay and professional development reforms that are supposed to increase and direct effort toward improving teachers instructional abilities while on the job. Additionally, performance pay and job-embedded professional development are supposed to raise human capital by attracting more capable individuals into the teaching profession and improving retention of good teachers (Podgursky & Springer, 2007). In turn, student learning will improve at a faster rate than would have occurred had performance pay and professional development not been reformed. A review by Podgursky and Springer (2007) found few rigorous studies of TISs and their impact on student achievement. Those that used a matched comparison group design found either positive or insignificant differences in test scores and other student outcomes relative to the comparison group, but the programs tended to be so

specific that the findings may not generalize to large-scale teacher performance pay programs, such as Q Comp and TAP.

The interaction between incentives and professional development can be complicated by multiple levels of implementation: federal, state, district, school, and grade-level teams. Q Comp exemplifies implementation divided along state and district lines. That is, Q Comp is a statewide TIS in which all schools are expected to participate if a district voluntarily applies and is accepted. This contrasts with Race to the Top, which is a federal program that expects all districts and schools to participate if a state voluntarily applies and is accepted.

Evaluating Q Comp

By what criteria and methods should Q Comp be evaluated? This evaluation study applies an *objectives-oriented* approach that incorporates statistical methods from a variety of disciplines in order to estimate Q Comp's impact on student achievement. Several factors warrant evaluating Q Comp in such a manner:

1. its scale of participation and expenditures
2. the relatively new and politically contentious educational policies it represents
3. voluntary participation at the district level
4. Q Comp's stated goal "to improve student learning" and (wide agreement that educational programs should ultimately impact student outcomes) makes student achievement an important criterion for judging its effectiveness.

Before describing the prior evaluation findings and the warrant for methods applied in this study, it is important to define evaluation and review alternative approaches.

Fitzpatrick, Sanders, and Worthen (2004) define program evaluation as "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit)" (p. 5). *The Program Evaluation Standards* (Yarbrough, D. B., & Joint Committee on Standards for Educational Evaluation, 2011) offer a more comprehensive definition of program evaluation:

> the systematic investigation of the quality of programs, projects, subprograms, subprojects, and/or any of their components or elements, together or singly, for the purpose of decision making, judgments, conclusions, findings, new knowledge, organizational development, and capacity building in response to the needs of identified stakeholders, leading to improvement and/or accountability in the users' programs and systems, ultimately contributing to organizational or social value. (p. XXV)

Evaluation can also be defined in terms of its relationship to scientific research. Evaluation and research are closely related and inform each other, but they differ in terms of disciplinary orientation, how findings are generalized, involvement of stakeholders, and professional standards. Table 2.1 specifies the characteristics that distinguish evaluation from research, according to Fitzpatrick et al. (2004) and Yarbrough and Joint Committee on Standards for Educational Evaluation (2011).

Table 2.1. Characteristics that distinguish evaluation from research

Evaluation	Research
• Interdisciplinary	• Intradisciplinary
• Pertains to a specific program, policy or project	• Pertains to scientific laws or theories about relationships
• Program stakeholders are the primary audience and source of evaluation questions	• Other researchers in the discipline are the primary audience and source of evaluation questions
• Describes and makes judgments	• Adds new knowledge and makes generalizable conclusions
• Held to standards for accuracy, feasibility, propriety, and utility	• Held to standards for internal and external validity
• Participants, policies and other factors influence sample and selection into the program	• Researcher controls sampling and treatment assignment

Within the field of evaluation, there are several different dimensions--purposes, questions, roles, and types of data--that an evaluator may emphasize (Fitzpatrick et al., 2004). Evaluations can serve a *formative* purpose, which typically occurs earlier in the life of a program and emphasizes diagnosing strengths and weaknesses to improve the program, or a *summative* purpose, which typically occurs after the program has matured and emphasizes judgments to inform major decisions about a program. Evaluation questions may focus on assessing a program's *needs*, monitoring *processes*, or inferring achieved *outcomes*. An evaluator may work *internally* (e.g., within the organization implementing a program), where the evaluator's familiarity with a program and rapport with stakeholders can strengthen the evaluation, or *externally* (e.g., in an evaluation consulting firm), where the evaluator may have more freedom to objectively evaluate a program. Evaluation data (and associated methods) can be qualitative (i.e., nonnumerical, narrative) or quantitative (i.e., numerical, statistical). The dimensions of evaluation are neither exhaustive nor mutually exclusive, meaning an evaluator can

emphasize opposite ends of a dimension and/or freely combine emphases from many different dimensions.

Fitzpatrick et al. (2004) have distilled evaluation dimensions, theories articulated by scholars, and actual practice in the field into an evaluation *orientation* continuum. On the *intuition-pluralist* end of the continuum lies *participant-oriented* evaluation. Rooted in subjectivist epistemology, it stresses interacting with program stakeholders to document and interpret their multiple perspectives/realities with the goal of understanding the program better (and not necessarily to support judgment). *Expertise-oriented* evaluation stresses the structured involvement of experts who bring knowledge and experience (and subjectivity) to judge a program's merit or worth (e.g., as part of a blue-ribbon panel). *Consumer-oriented* evaluation stresses proactively providing evaluation information to consumers, typically in the form of criterion checklists, so that they can make informed selections. *Management-oriented* evaluation stresses providing evaluative information to program managers in a manner that supports rational decision making. *Objectives-oriented* evaluation lies on the *utilitarian* end of the continuum. Rooted in objectivist epistemology, it stresses operationalizing and objectively measuring program objectives over time to determine the extent to which processes and/or outcomes were impacted.

This evaluation study applies an objectives-oriented approach. It incorporates statistical methods from a variety of disciplines in order to estimate Q Comp's impact on student achievement. Intuition-pluralist approaches and qualitative methods are beyond the scope of this study, but it features a value-added analysis intended to makes those approaches more feasible in the future. This study qualifies as an evaluation because it is

interdisciplinary, pertains to a specific program, and because sampling and treatment assignment were not controlled (see Table 2.1). However, it combines some elements of evaluation and scientific research. That is, although this study focuses on Q Comp and evaluating its student achievement objective, the questions it addresses and the novelty of Q Comp's policies make it likely that researchers and others beyond Minnesota will take interest in the findings and attempt to generalize them. This necessitates recognizing researchers as an audience, in addition to local stakeholders. As such, this study prioritizes addressing validity threats (as a research study might) and attending to the *accuracy* Program Evaluation Standards (Yarbrough, D. B., & Joint Committee on Standards for Educational Evaluation, 2011).

This evaluation also qualifies as a meta-evaluation of SRD methodology. *The Program Evaluation Standards* (Yarbrough, D. B., & Joint Committee on Standards for Educational Evaluation, 2011) define metaevaluation as "systematic evaluation of evaluations and their subcomponents" (p. 227). SRD is a major subcomponent of this evaluation, and it is a relatively new method, especially in the evaluation of educational objectives. The criteria for judging SRD are pre-test balance and fit and post-test distance coefficients.

The next section reviews findings from four prior evaluations of Q Comp, some of which were introduced above. Two evaluations were conducted early in the program's life. Their purposes were largely formative and management-oriented. One was management-oriented with respect to MDE and the other with respect to the Minnesota Legislature. Both involved stakeholders in data collection, and they mixed qualitative and quantitative data and methods in order to evaluate Q Comp in terms of process

changes. They did not evaluate Q Comp in terms of student achievement objectives, saying the program's short life and voluntary nature prevented them from doing so defensibly. At the time this thesis was started, there had been no evaluation of Q Comp in terms of it's impact on student achievement--a gap I aimed to fill by applying quasi-experimental statistical methods. Since then, two evaluations of Q Comp's impact on student achievement have been conducted. They were more summative in purpose than the two earliest evaluations, and one involved documenting stakeholders' multiple perspectives/realities in order to advance understanding of Q Comp.

Prior evaluations of Q Comp's impact on practice and student achievement

To what degree has Q Comp changed incentive and professional development practices? Hezel Associates (2009) and the Office of the Legislative Auditor (2009) conducted one-time interviews and surveys asking about perceptions and implementation early in the program's existence. Both found mixed results. About half of respondents reported no change in the career advancement and job-embedded professional development components of Q Comp (i.e., either due to prior practice or participation in one of the pilots that began 2003 or due to no discernible implementation). About half also reported confusion about what constitutes performance (as opposed to effort). Even though moving forward in steps based on performance is not supposed to be automatic, evidence of low standards for judging performance and inadequately alternative schedules in some districts and charter schools led the Office of the Legislative Auditor (2009) to conclude, "It is not clear how much of a change these reformed salary schedules represent" (p. 5). Some Q Comp districts and charter schools simply replaced experience steps with performance level steps along rows in the matrix; others replaced

their schedule with a formula that adds pay for simply taking on additional responsibilities. Less than half of teachers agreed with the statement, "the Q Comp program has improved classroom teaching at my school," and only a third agreed with the statement, "the Q Comp program will lead to increases in students' performance on standardized tests at my school."

Schwartz (2012) and Sojourner et al. (in press) found somewhat stronger evidence of shifts in incentives and professional development among Q Comp participants. Teachers and principals in 55 randomly sampled districts took the National Center for Educational Statistics' Schools and Staffing Survey before Q Comp (in 2003-2004) and again after they started participating (in 2007-2008). Schwartz (2012) examined responses of repeatedly surveyed sites and found that Q Comp teachers reported a significant shift toward performance pay by about $750 or 1% of their total pay. Teachers also reported about a 50% increase in some types of professional development compared to respondents who had not participated in Q Comp. Sojourner et al. found that no Q Comp district reported paying for performance before participating and 58 percent did afterward, which compares with four percent of non-participants. A phone survey of district human resource professionals by Sojourner et al. (in press), which did not even mention Q Comp, found that about 90 percent of participants reported paying for student outcomes and classroom observations compared to none of the non-participants surveyed. However, when they asked about paying for years of experience and levels of education, about 95 percent of Q Comp participants and 100 percent of non-participants responded affirmatively, which suggests that the alternative salary schedules

required by Q Comp represent supplements to than a replacements of the traditional schedule.

Schwartz (2012) conducted semi-structured interviews of Q Comp designers and leaders at the state and local levels to qualitatively fill in additional details about Q Comp's implementation. The state-level interviews revealed Q Comp did not end up mirroring TAP to the degree that the designers intended. Statutory language and local educational control hamstrung MDE's authority to impose strong performance pay and make improvements over time. "[M]ost districts hoped to spread the money to as many teachers as possible while the state pushed for more drastic versions of merit pay" (Schwartz, 2012, p. 82). Additionally, MDE lacked the capacity to implement the instructional improvement components of Q Comp. Q Comp legislation was passed 17 days after the law's stated start date and provided no funding to MDE to implement the program. MDE did what it could by shifting already employed staff, interpreting the law, articulating rules, and ensuring compliance. Implementing the instructional improvement component was left to local school districts, partly due to MDE's limited capacity in that area and partly by design in order to encourage innovative instructional improvements. The state teachers union initially filled the gap left by MDE in terms of providing statewide support for professional development reform, but conflicts with MDE and a change in leadership led the union to back away from that role.

For the local-level interviews, Schwartz (2012) identified key informants at three school districts that possessed similar student characteristics but were known to MDE and Education Minnesota (the state teacher's union) to differ in their approaches to teacher professional development. The districts were Farmington, Centennial and St. Francis,

and the interviews provide a typology for understanding how districts implemented the instructional support components of Q Comp.

1. Farmington emphasized *peer teachers* as the source of new instructional knowledge.
2. Centennial emphasized PLCs in which teachers reviewed *student data* together.
3. St. Francis emphasized *external sources of new knowledge* (e.g., the American Federation of Teachers' Education Research and Dissemination Program).

As a result of political forces and MDE's limited capacity, Q Comp became a TIS with limited but well-documented incentive strategies applied at different levels and highly variable but poorly documented instructional improvement strategies determined at local levels. Sojourner et al. (in press) call Q Comp "pay-for-performance-centered human-resource management reform" (P4P-centered HRM). The terms "centered" and "reform" are qualifiers that allude to uncertain fidelity to pay-for-performance (i.e., to an ideal type or variations found in the private sector) while also acknowledging evidence of true and long-term reform *as indicated by* paying for performance *and* professional development. "P4P-centered TIS" might be a fitting label for the universe of generalization from which Q Comp and its strategies were "sampled" (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001).

To what degree has Q Comp impacted student achievement? Early evaluations of Q Comp by the Office of the Legislative Auditor (2009) and Hezel Associates (2009) stressed the importance of evaluating Q Comp in terms of its influence on student achievement. However, the Office of the Legislative Auditor concluded that the voluntary nature of Q Comp and its short existence made estimating its influence on

student achievement too difficult. Hezel Associates, noting the same limitations, estimated correlations between the number of years of Q Comp participation and scores on the Minnesota Comprehensive Assessments (MCA-II) in 2008. The correlations were significantly greater than zero but small in size, ranging from 0.015 to 0.093.

Schwartz (2012) overcame some of the limitations encountered in the early evaluations of Q Comp's influence on student achievement by longitudinally analyzing school-level achievement outcomes from 2004-2010 in participating and non-participating districts. He regressed school means of student z-scores for math and reading Minnesota Comprehensive Assessments (MCAs) on Q Comp participation and control variables and pre-test trends while weighting by the count of students. He found that Q Comp did not significantly influence math achievement (d = 0.021, se = 0.015) or reading achievement (d = 0.010, se = 0.011). The findings were robust to different specifications and to substituting school means of MCA scores with student-level scores from the Measures of Academic Progress (MAP) test, which a subset of districts and schools administered voluntarily. Even though he did not find a significant change in mean achievement, a subsequent analysis suggested that variation in student achievement decreased/narrowed, which Camerer and Hogarth (1999) have noted is a common result of incentive programs.

Sojourner et al. (in press) also conducted a regression analysis of math and reading achievement from before and after Q Comp began in participating and non-participating districts while controlling for student socioeconomic characteristics. Their dependent variables were *students' test scores* (i.e., not school-level means) from 2004-2010. They conclude that Q Comp significantly impacted reading achievement (d =

0.031, *se* = 0.015) but not math achievement (*d* = 0.004, *se* = 0.021) as measured by the state-required MCAs. Adding indicators for years of participation, they found evidence that Q Comp's impact on reading achievement grew larger with duration of participation. Findings were similar for the subsample of districts that voluntarily administered the MAP test and held up under different model specifications and falsification tests.[1]

In economic terms, Sojourner et al. (in press) estimate that Q Comp's reading achievement benefits exceed the program's cost by about 5 to 1. They base their benefit ratio on studies by Krueger (2003), Hanushek (2010), and Chetty et al. (2011) that suggest a social return of about a million dollars per standard deviation increase in student achievement. For Q Comp, their point estimate of 3% of a standard deviation increase in reading translates into a benefit of $30,000 relative to an estimated cost of $6,500 per classroom per year (i.e., 25 students at $260). They attribute effectiveness to the long-term nature of district policy changes under Q Comp and increased productivity of less experienced teachers (as opposed to changes in expenditures, turnover, experience, and effort among senior teachers).

Taken together, prior evaluations of Q Comp reveal mixed results regarding Q Comp's impact on practice and student achievement in the first four years of its existence. They establish that Q Comp has successfully overcome a number of difficulties to implement enduring changes to teacher pay and professional development but in ways that are difficult to characterize and measure. The program could have faltered under the

1 This study began around the same time as Schwartz's (2012) and Sojourner et al.'s (in press) studies of Q Comp. The methods and model specifications applied in this study were chosen without knowledge of the other studies' methods and specifications, even though regression methods from all three converge with regards to estimating Q Comp's impact relative to both pre-Q Comp participation and non-participating districts.

weight of competing political demands, especially given MDE's lack of funding and capacity, but program documentation, surveys and interviews indicate that Q Comp has struck a balance between performance pay advocated by incentive-focused designers and job-embedded professional development favored by teachers. And it has managed to attract a growing number of districts without surpassing budget allocations. Despite evidence of sustained reform, evidence of the degree to which Q Comp has impacted practice and student achievement remains mixed due to methodological challenges. That is, the program's voluntary nature, latitude for variable pay and locally implemented professional development practices introduce selection/history and measurement error, which threaten the validity of conclusions about Q Comp and the ability to identify and scale up best practices. One quasi-experimental study found evidence of Q Comp's impact on student achievement lacking; another found a small reading effect size that nevertheless suggests potentially large economic benefits. As Q Comp enters its ninth year, further study is necessary to measure its overall impact and which components, combinations, and intensity of components (observations, PLCs, career advancement, student learning goal setting, performance pay, and/or alternative salary schedules) were most effective at accelerating student achievement.

Spatially enabled evaluation

Educational evaluators are overlooking an important source of information in their endeavor to improve educational outcomes: *space*. Unlike *time*, which is unidimensional and an indispensable reference in evaluation research, spatial/geographic information is multidimensional and underutilized. In particular, geographic information systems (GIS) and spatial statistical analysis are underutilized in applied research when

compared to research in other fields (Renger et al., 2002; Tate, 2008). Research in public health, political science, and economics has become *spatially enabled*, meaning spatial methods are used regularly to form research questions, sample and collect data, analyze data, and disseminate results (e.g., in the form of maps; Waller & Gotway, 2004).

Spatial regression discontinuity (SRD) design represents a facet of *spatially enabled evaluation*. Patton (1997) defines evaluation as "systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming." Spatially enabled evaluation is evaluation, policy analysis, or applied research made possible with spatially referenced data. Just as longitudinal data have been referenced to points in time, spatial data have been referenced to points in space, such as latitude and longitude coordinates. The term *spatial* is broad and encompassing of *geographic*. *Geographic information system* (GIS) refers to rapidly expanding computer technology that handles spatial data and is often used as a synonym for the spatial information it handles (Graham, 2001).

Spatial information

GIS handles four main types of purely spatial information defined by geography and lacking attributes/substance, such as demographic information (Bivand, Pebesma, & Gómez-Rubio, 2008; Banerjee, Carlin, & Gelfand, 2004; Ormsby et al., 2001).

1. A *point* is a single location, such as a global positioning system (GPS) satellite reading or a street address pinpointed (i.e., *geocoded*) to a unique location.
2. A *line* is a series of straight line segments that connect a set of ordered points.

3. A *polygon* is an area enclosed by a set of lines, possibly containing holes (e.g., a polygon in the shape of a donut); also described as *areal*.
4. A *grid* is a collection of points or rectangular areas organized in a regular fashion; also described as *raster* or *lattice*.

Purely spatial information alone (i.e., coordinates) holds little value for evaluators and educational researchers, but GIS can enable evaluators to transform non-spatial, program-relevant information into spatial information. GIS technology can act as a relational database and join non-spatial data to purely spatial information when two sets of data share a common index (Renger et al, 2002). From a geographic perspective, joining data in this manner is equivalent to assigning explicit attributes to spatial features (e.g., assigning student proficiency rates to school points; Bivand, Pebesma, & Gómez-Rubio, 2008). From an evaluation perspective, it is referred to as *spatially referencing* non-spatial information (e.g., joining individual students to school district polygons).

Mapping

Spatial methods involve spatially referenced information either as a key component of research design or as an object of statistical analysis. Visualizing spatially referenced data in the form of maps is a basic and widely used method in spatially enabled evaluation. Maps can benefit both evaluators and consumers of evaluation information, but maps also carry risks. Maps represent an alternative and complement to stories, graphs, tables, and figures traditionally used to engage stakeholders and present evaluation information (Renger et al., 2002). Maps may prove helpful during the divergent phase of evaluation planning, when it is important to engage stakeholders in order to generate key evaluative questions (Talen and Shah, 2007; Fitzpatrick et al.,

2004). Maps can also promote comprehension of findings during the dissemination phase of an evaluation by acting as *adjunct aids* to text in an evaluation report, but only if the maps are of high quality and produced from the findings with fidelity (Carney & Levin, 2002; Verdi & Kulhavy, 2002). Visualization alone is insufficient for systematically judging the merits of a program, so maps created for evaluation purposes should depict measures of change in key outcomes over time, when possible, in order to facilitate judgment (Renger et al., 2002; Brown, 2005). The risks of mapping evaluation data may outweigh the potential benefits given that flat geographic maps automatically distort the earth's three-dimensional surface (Monmonier, 1996), that evaluators routinely face political and ethical pressures (Fitzpatrick et al., 2004), and that maps can inadvertently reveal where program participants live (Banerjee, Carlin, & Gelfand, 2004).

Spatially enabled research design

Spatially referenced data can serve as a key element of research design. Talen and Shah (2007) developed survey instruments that prominently featured GIS-produced maps. Brown (2005) illustrates how spatial information can facilitate stratified random sampling and power analysis for survey implementation. He also notes that GIS can help evaluators conduct cluster randomized trials—an experimental design being promoted by the Institute of Education Sciences (2007) of the U.S. Department of Education. Spatial data can also support quasi-experimental evaluations that allow causal conclusions. Propensity score matching with information about program participants' locations can reduce selection bias and improve causal estimates (Cook et al., 2008). Holmes (1998) evaluated state-level business policies using *spatial regression discontinuity* analysis,

treating state borders as sharp cutoff points and estimating the effect of manufacturing regulation at the borders.

Spatial analysis

Unlike time, which follows a single ordered dimension in one direction (past ⇉ future), space is multidimensional and multidirectional (north ⇆ south, east ⇆ west; Diggle, 2004). Data referenced simultaneously to both time and space (i.e., spatiotemporal data) occupy three or more dimensions. According to Diggle (2004), spatial statistical analysis and the problem of spatially dependent observations motivated Fisher to advocate for randomization and blocking in agricultural field trials and beyond. Diggle defines spatial statistics as "the formulation and analysis of stochastic processes indexed by (typically two-dimensional) space rather than by one-dimensional time, in which context the increase in dimensionality is less important than the loss of a natural ordering to the index set" (p. 702). That is, Y_t can depend on Y_{t-1}, Y_{t-2} and earlier periods but not on Y_{t+1} or later periods, whereas Y_s can depend on Y_{s-1}, Y_{s-2}, Y_{s+1}, Y_{s+2}, and more distal points.

Estimating and accounting for spatial correlation is an important step in spatial statistical analysis. The "first law of geography," according to Tobler (1970), is "everything is related to everything else, but near things are more related than distant things" (p. 236). Spatially correlated observations violate the statistical assumption of independent observations and reduce statistical power (Waller & Gotway, 2004). An educational consequence of Tobler's law is that spatial context may be an important and substantively interesting factor to consider in educational research and decision making. For example, where a student lives continues to determine school choices and quality

(Orfield & Wallace, 2007), and neighborhood conditions help explain disparate educational and health outcomes (Leventhal & Brooks-Gunn, 2000). Statisticians have developed practical approaches for dealing with spatial correlation. The choice among approaches usually depends on whether spatial correlation is considered substantively interesting or a nuisance (Anselin et al., 1996). If spatial correlation is substantively interesting, then distance or spatially lagged dependent variables can be treated as explanatory variables. If spatial correlation is considered a nuisance, then multilevel modeling (a.k.a. hierarchical linear modeling) or geostatistical modeling would be appropriate.

Generalized causal theory and methods

Shadish, Cook, and Campbell (2002) define internal validity as the "validity of inferences about whether observed co-variation between *A* (the presumed treatment) and *B* (the presumed outcome) reflects a causal relationship from *A* to *B* as those variables were manipulated or measured" (p. 53). A causal claim must meet three conditions in order to be internally valid:

1. *precedence*: the theorized cause must precede the observed effect
2. *relationship*: the cause and effect must be related
3. *no competing explanation*: competing explanations, such as a confounding variable, cannot falsify the inference.

If precedence of the cause is in doubt, if the relationship between cause and effect is not conclusive, or if probable alternative/competing causes cannot be ruled out, then a strong causal claim is not warranted. Cross-sectional studies that lack a pre-test observation may not meet the precedence condition; studies that take multiple observations over time

are better equipped to meet the precedence condition. Even if the precedence condition is met, studies that lack statistical power, base conclusions on unreliable measures of cause and/or effect, or suffer from other threats to statistical conclusions cannot claim to be internally valid. If the first two conditions are met, but participants in a study are allowed to self-select into the treatment condition, the treatment group experiences an event that coincides with the start of the treatment, or any other factor may be responsible for the observed relationship, then competing explanations threaten the internal validity of the study.

Shadish et al. (2002) go on to define external validity as "the extent to which a causal relationship holds over variations in persons, settings, treatment variables, and outcome." External validity concerns the generalization of internally valid inferences to variations both within and beyond an experiment. Generalizations may be made from narrow (the persons, settings, treatments, outcomes, and/or variables involved in the experiment) to broader situations, from broad to narrow, at similar levels (e.g., from one location to a nearby location) to similar or different groups, or from a random sample to a population. The authors acknowledge there are risks of generalizing beyond an experiment's particulars, but they contend that generalization is just as important as establishing internal validity because science requires incremental extensions of both theory and experimentation into new realms. Establishing external validity involves using design and/or analysis to rule out the possibility that causal relationships depend on variations in persons, settings, treatments, and outcomes. Testing the null hypotheses of no moderation/interaction between the treatment and other variables is an analytic approach to establishing external validity.

Randomized experiments

A randomized experiment is considered the *simplest* method for inferring causality because the experimenter prospectively exercises a high degree of control over the timing of measurement and over potentially confounding variables (Holland, 1986). Experimenters generally take measures before and after treatment, and by definition, they randomly assign experimental units to treatment conditions in order to evenly balance potentially confounding variables between groups. Prospective random assignment makes balance highly probable, but balance should still be checked (Shadish et al., 2002).

Rubin's (1974) *potential outcomes model* highlights the benefits of balance. It encompasses random experiments and allows competing causal explanations to be ruled out when random assignment fails or is not possible. In the case of a simple experiment, the potential outcomes model states that since an experimental unit *u* has the potential of being in either the treatment or control group, but it is impossible to observe the true within-unit causal effect $Y_t(u)-Y_c(u)$ because a unit cannot be exposed to the treatment and control condition simultaneously (Holland, 1986). That is, observations of $Y_c(u)$ are missing for those in the treatment group, and conversely, $Y_t(u)$ observations are missing from the control group. The next best alternative is to calculate the *average treatment effect* (ATE) $\bar{Y}_t-\bar{Y}_c$ from observed outcomes and assume that one group's observed outcomes accurately represent the other group's unobserved outcomes, or *counterfactuals*. Informally, the potential outcomes model requires experimental units in one group to be just like those in the other group, except for the treatment condition. Formally, it requires the distributions of potential outcomes to be the same, independent

of treatment assignment but conditional on confounding covariates: $Y_c, Y_t \perp T | X$ (Gelman & Hill, 2007).

Ignorability refers to the ability to disregard participants' characteristics and how they were assigned to a treatment condition when estimating causal effects under the potential outcomes model. Random assignment during the design phase supports ignorability and simplifies estimating causal effects because participants' pre-existing characteristics and preferences do not (presumably) influence their probability of being in a given treatment condition. Random assignment, balance and ignorability can be tested by statistically regressing actual treatment condition on a set of *k* variables to test the null hypothesis of independence: $H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$. If random assignment was not implemented and/or the null hypothesis is rejected, then the mechanism and covariates cannot be ignored (i.e., a more complex causal analysis is required); failure to reject would verify random assignment and justify simply ignoring the assignment mechanism and covariates during the causal analysis phase.

Quasi-experiments

Quasi-experiments are "almost-like" randomized experiments (Cook, 1991). Both aim to infer causal connections by implementing research designs that emphasize the timing of observations (pre- and post-test) and compare groups (treatment and control). The key difference is that quasi-experimentation is weaker because research participants are not randomly assigned to the treatment group. As such, quasi-experiments require researchers to closely attend to the *nonignorable* assignment mechanism and its consequences for internal validity. That is, the causal analysis must condition on

covariates in order to invoke independence of potential outcomes with respect to the treatment. Sometimes nonignorable selection is justifiably ignored when selection/noncompliance is an expected feature of a real-world program (i.e., when the intent-to-treat effect is more appropriate than the treatment-on-treated for calculating the return on investment in a publicly funded program).

Cook (1991) asserts that evaluators should strive to choose designs that give the most complete information about selection into the treatment group. Randomized experiments generally provide the most information, followed by regression discontinuity designs. If those designs are not feasible, then evaluators should choose designs that offer the most pre-treatment information for assessing selection and maturation. Such designs include (from most to least information):

- interrupted time series, which involves repeatedly measuring a participant's outcomes both before and after introducing a treatment and regressing them on time in order to estimate how the treatment impacted level and/or slope
- double pre-test, which involves repeated pre-test measures but only one post-test measurement
- equivalent comparison group with single pre-test, which involves establishing a comparison group that is ignorably different at pre-test but without random assignment
- nonequivalent comparison group with single pre-test, which involves statistical adjustments for nonignorable differences at pre-test.

Researchers can improve the validity of quasi-experiments by: 1) introducing and re-introducing treatments to see if a pattern is mirrored in the outcome; 2) making multiple pre-test observations to establish pre-treatment patterns over time; 3) forming multiple nonequivalent comparison groups representing different levels of pre-test performance; and 4) administering different levels of treatment exposure in accordance with expected effect sizes.

Causal inference through prospective research design is preferable over statistical adjustments to retrospectively observed data (Cook, 1991). Nevertheless, statistical controls are important when an evaluator cannot exert control over research design. One of the largest potential benefits of spatially enabled evaluation is the large amount of control variables made available by spatially joining community-level attributes to individuals (Renger, Cimetta, Pettygrove, & Rogan, 2002; Quon Huber, Van Egeren, Pierce, & Foster-Fishman, 2009; Graham, 2001).

Regression discontinuity designs

Spatial regression discontinuity (SRD) design entails estimating a geographically local average treatment effect (GLATE) by statistically regressing an outcome on distance from a border that was used to assign persons to the treatment condition. SRD design and analysis incorporate elements of traditionally univariate RD, multivariate RD, and spatial analysis, but they differ importantly. SRD designs differ from traditionally univariate RD designs because multiple variables (i.e., longitude and latitude) determine treatment assignment. SRD designs differ from other multivariate RD designs because the assignment variables are not inherently ordered. SRD analysis differs from spatial

analysis when order is imposed by calculating distance. Before describing SRD designs, it is important to review traditionally univariate RD designs and multivariate RD designs.

Univariate regression discontinuity design

Regression discontinuity is a quasi-experimental design that assigns participants to a treatment (e.g., tutoring) based on a sharp cutoff point (e.g., poverty criterion) along a continuous assignment variable (e.g., income), allowing a local average treatment effect (LATE) to be estimated at the cutoff (Imbens & Lemieux, 2007). Persons near each other on both sides of the cutoff are assumed to be just like each other except for treatment assignment. Piecewise regression analysis (a.k.a., broken stick, segmented, or switching regression) is used to estimate LATEs at the cutoff by estimating intercepts and slopes on both sides of the cutoff.

Prospective regression discontinuity designs, whereby the researcher can both assign participants and enforce treatment, are ideally conducted as follows:

1. take a pre-test measure of the assignment variable (denoted as X_i)
2. establish a cutoff point somewhere along the pre-test variable (X_c), preferably at the mean to strengthen statistical power
3. assign experimental units with pre-test values on one side of the cutoff to the treatment condition and those on the other side to the control condition
4. take outcome measures after treatment
5. use exploratory data analysis to specify a plausible functional form for the relationship between the assignment and outcome

6. statistically regress the outcome measure Y_i on the pre-test variable centered at zero $(X_i - X_c)$ interacted with a dummy variable representing the treatment condition (e.g., Z_i = 1 if $X_i - X_c$ < 0; 0 otherwise)
7. examine the statistical significance and size of the coefficient for the treatment dummy variable Z_i, which is an estimate of the true LATE.

Figure 2.3 shows four possible results of the aforementioned design, with assignment and outcome variables originally distributed as T ~ N(50, 100) and the true models defined as:

Model 0: $Outcome_i = 50 + 0.5\,X_i + \varepsilon_i$

Model 1: $Outcome_i = 50 + 10\,Z_i + 0.5\,X_i + \varepsilon_i$

Model 2: $Outcome_i = 50 + 0.5\,X_i - 0.5\,Z_i\,X_i + \varepsilon_i$

Model 3: $Outcome_i = 50 + 10\,Z_i + 0.5\,X_i - 0.5\,Z_i\,X_i + \varepsilon_i$.

Model 0 (top left) represents no effect: the RD cutoff influences neither the level nor the slope. It also represents what a researcher would ideally see before treatment implementation (i.e., pre-test measures). A multiple-pre-test analysis is very helpful for determining the functional form beforehand, which helps rule out validity threats that stem from incorrectly specifying the functional form between the assignment and outcome. Models 1 and 3 (top right and bottom right) represent large effect sizes at the cutoff: treatment assignment changes the conditional mean at the cutoff. Local intercept effects are typically the focus of RD studies rather than slope changes for reasons related to external validity discussed below. The plot of Model 2 (bottom left) shows no effect at the cutoff but a non-local effect because the RD cutoff only influences the slope. A

change in slope but not intercept might occur if a treatment is best suited to participants far from the cutoff, such as remedial instruction when the cutoff is made near the average of prior test scores.

Figure 2.3. Outcomes from a regression discontinuity design: Local average treatment effects and interactions between assignment and treatment

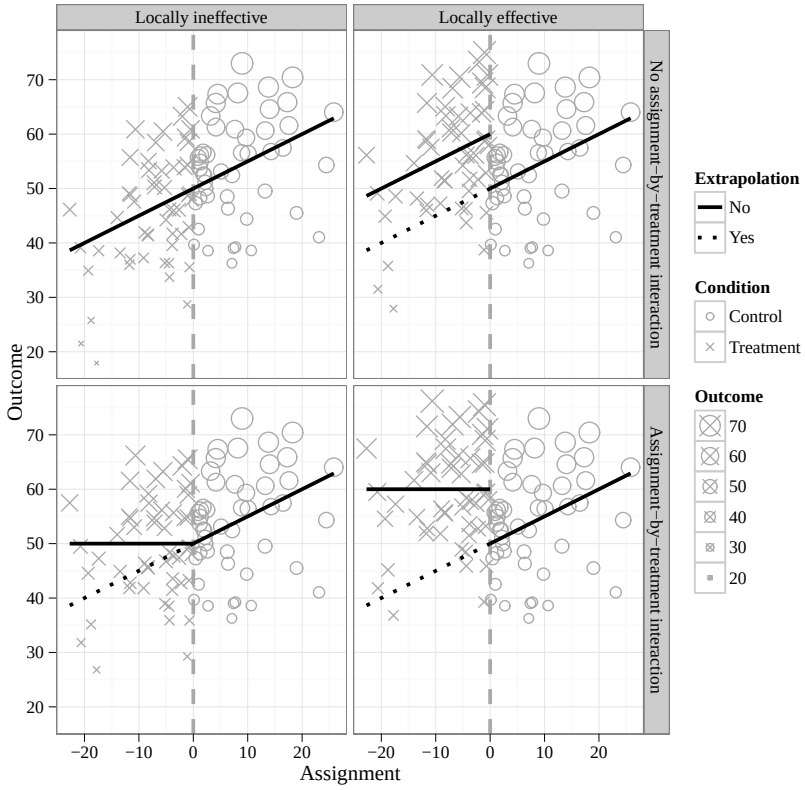


Figure 2.3 also illustrates the interplay between internal and external validity within a RD study. As noted by Shadish et al. (2002), external validity often concerns generalization to persons, settings, treatments, and outcomes beyond a study, but it also involves generalizations within a study. The primary goal of a RD design is to locally estimate effects (i.e., to estimate LATEs), unlike a randomized controlled trial in which the goal is to simply estimate a non-local average treatment effect (ATE) thanks to

ignorability. RD designs involve a tradeoff: internally valid estimates in exchange for estimates that do not generalize beyond the cutoff (Campbell, 1969). Generalizations beyond a cutoff may be warranted if multiple pre-test observations were made, allowing the pre-test fitted line to be estimated without bias and with precision. That is, a researcher may be able to use past data to extrapolate the control group's fitted line beyond the cutoff and use it as a counterfactual to estimate non-local average treatment effects. Extrapolating the fitted line for the control group in the Model 1 plot (top right) suggests that the large effect remains constant over the treatment range of the assignment variable (i.e., the local effect is externally valid within the study). The Model 2 plot (bottom left) shows no effect locally, but it suggests that the treatment may become increasingly effective for those scoring far lower than the treatment cut score. The Model 3 plot (bottom right) suggests the large local effect becomes increasingly effective for those with the lower assignment scores.

Incorrectly specifying the functional form for the regression of the outcome on the assignment can bias estimates and threaten the internal and external validity of inferences from a RD study. Figure 2.4 illustrates incorrect specifications. The dark lines represent truly cubic relationships, and the lighter lines show the fitted lines estimated by ordinary least squares with a linear specification. The true models are defined as:

Model 0: $Outcome_i = 50 + 0.5\,X_i - 0.025\,X_i^2 + 0.0025\,X_i^3 + \varepsilon_i$

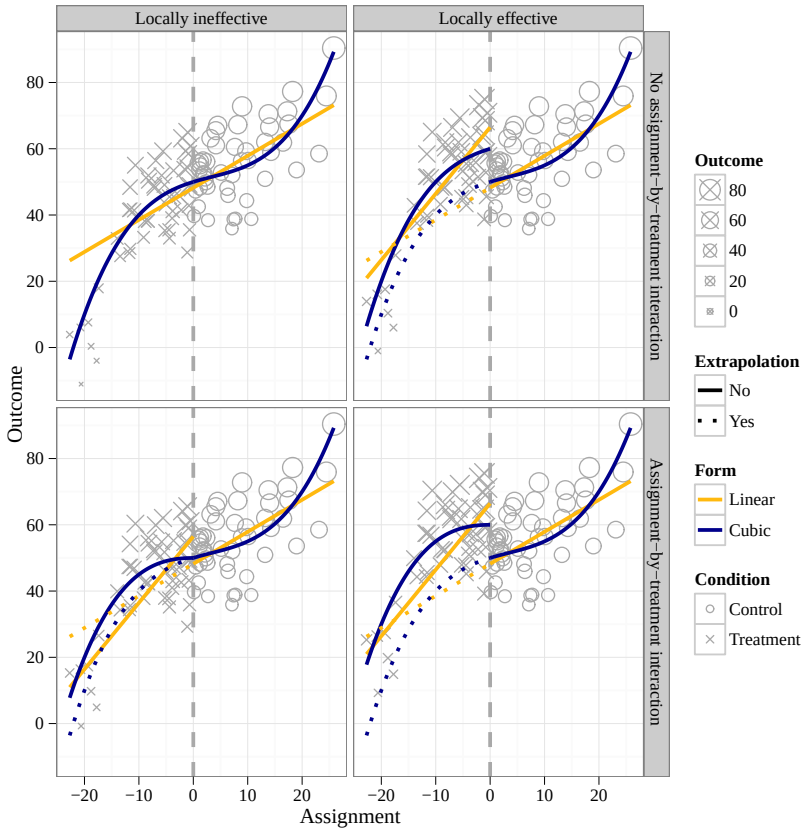
Model 1: $Outcome_i = 50 + 10\,Z_i + 0.5\,X_i - 0.025\,X_i^2 + 0.0025\,X_i^3 + \varepsilon_i$

Model 2: $Outcome_i = 50 + 0.5\,X_i - 0.025\,X_i^2 + 0.0025\,X_i^3 - 0.5\,Z_i\,X_i + \varepsilon_i$

Model 3: $Outcome_i = 50 + 10\,Z_i + 0.5\,X_i - 0.025\,X_i^2 + 0.0025\,X_i^3 - 0.5\,Z_i\,X_i + \varepsilon_i$.

Observations in the plot of Model 0 (top left of Figure 2.4) can be thought of as pre-test observations. Nonparametric smoothing of those observations would have revealed the appropriateness of a cubic specification. The linear specifications overestimate the true LATEs when they are greater than zero. That is, in the plots of Model 1 (top left), 2 (bottom left) and 3 (bottom right) the "breaks" between the solid lighter lines at the cutoff exceed the breaks between solid darker lines by several points. The linear specification also yields biased estimates beyond the cutoff (see dashed extrapolation lines in the treatment regions). The bias illustrated by the misspecification of the true functional form could have large consequences in the real world if the observed/estimated relationships were taken to be true and used to make decisions (e.g., regarding dosage), especially decisions that over-generalize beyond the cutoff.

Figure 2.4. Consequences of misspecification



Regression discontinuity is favored by federal agencies and evaluation theorists because well-designed regression discontinuity studies yield causal estimates that are comparable to those derived from randomized controlled trials (U.S. Department of Education, 2005; Cook, 1991; Cook et al., 2008). The treatment assignment process is completely known and perfectly measured—a feature that regression discontinuity shares with randomized controlled trials (Shadish et al., 2002). However, in contrast to a randomized experiment in which the assignment variable is assumed to be independent of all other variables, the regression discontinuity treatment assignment mechanism is *determined* by a continuous pre-test variable (Imbens & Lemieux, 2007). In terms of

methodological rigor (for ruling out validity threats), simplicity, statistical power, and opportunities for application, RD designs fall between randomized experiments and retrospective observational studies. Randomized experiments offer the most rigor and simplicity and power, but they are often not feasible. Observational studies are highly feasible, but they rarely warrant causal conclusions because the assignment mechanism is unknown and may require a large number of control variables.

Prospectively designed and controlled RD studies are superior to RD natural experiments. Cook (2008) notes that RD designs are often more feasible than randomized experiments because RD "does not require researchers or their proxies to directly manipulate the independent [assignment] variable" (p. 643). Evaluands (i.e., programs to be evaluated) are often charged with serving those most in need and with limited funding, which leads them to routinely set and enforce eligibility cutoffs. As such, RD natural experiments abound. Even if a researcher prospectively designs a randomized experiment, gaining rapport with the program and its participants may be easier with a RD design because cutoff-based assignments are so prevalent in society (e.g., income requirements for government assistance). A cost of choosing a RD design over a randomized experiment is loss of statistical power arising from collinearity between the assignment and treatment (Goldberger, 1972). Cappelleri, Darlington, and Trochim (1994) found that RD designs require between 2.34 and 2.73 times the number of participants as randomized experiments to detect a small and large effect size, respectively, with 0.8 statistical power. A major advantage of prospectively designed regression discontinuity studies is that a researcher can maximize efficiency by setting the cutoff at the mean of the assignment variable so that a large number of counterfactual

participants are located near the cutoff. RD natural experiments will generally offer less statistical power because evaluands set cutoffs to meet their needs and not with statistical power in mind.

Other advantages of prospectively designing and controlling regression discontinuity studies is that a researcher can 1) limit instances of "fuzzy" regression discontinuity (Campbell, 1969) and 2) rule out history validity threats. Fuzzy RD occurs when a treatment is misallocated to participants on the control side of the cutoff or vice versa. Fuzzy RD violates the assumptions of a sharp, deterministic assignment mechanism, although the *probability of receiving treatment* will be discontinuous to the extent that the cutoff was enforced (Hahn, Todd, & Van der Klaauw, 2001). Fuzzy RD and history are especially threatening to validity if the cutoff is well-established and well-known (Lee & Lemieux, 2010). If the cutoff is known, purposeful misallocation can result from political patronage (Campbell, 1969), good intentions/professional discretion (e.g., a program administrator who feels a participant is deserving of a service even though they do not technically qualify; Shadish et al., 2002; Urquiola & Verhoogen, 2009), or manipulation of the assignment variable by a program administrator or participant (McCrary, 2008). Manipulation is especially concerning because it would be hard to detect. That is, the degree of fuzziness cannot be determined and design or analytic remedies cannot be applied without the ability to identify fuzzy cases. Misallocation can also occur because the cutoff is simply unknown to those in charge of treatment assignment (e.g., due to poor communication). History "refers to all events that occur between the beginning of the treatment and the post-test that could have produced the observed outcome in the absence of that treatment" (Shadish et al., 2002; p. 56). If a

cutoff is well-established, such as a statewide cut score for classifying students as "proficient", then it is plausible that the cutoff will be used to determine a person's eligibility for several different programs (e.g., a tutoring program and a separate program that provides free books). Fuzzy RD and history threaten regression discontinuity natural experiments and weakly controlled RD designs. Fuzzy RD and history threats can be minimized by prospectively establishing a new cutoff that is unfamiliar to program administrators and participants and by monitoring and enforcing cutoff-based assignments. Analytic enhancements may help mitigate fuzzy RD or history validity threats (Imbens & Lemieux, 2007), but ruling out such threats through prospective design is preferable to statistical adjustments (Cook, 1991).

Multivariate regression discontinuity design

RD can be used to estimate LATEs along a multidimensional frontier or multiple cutoffs when more than one assignment variable determines treatment assignment (Wong, Steiner, & Cook, 2010; Papay, Willet, & Murnane, 2011). Wong, Steiner, and Cook (2010) refer to RD with multiple assignment variables as *multivariate regression discontinuity* (MRD) design. (Their use of the term "multivariate" refers to multiple assignment variables and not necessarily to multivariate analysis of multiple dependent variables.) They illustrate MRD by describing a scenario in which a student receives an educational intervention if either their math test score falls below a math cut score or their reading score falls below a reading cut score. In that design, the intervention is the same for all students regardless of the subject(s) in which they scored below the cutoff. The design yields two subject-specific frontiers, both of which represent subsets of univariate cutoffs if left unrestricted:

1. a math frontier along the reading score continuum
2. a reading frontier along the math score continuum.

A multidimensional frontier results from the union of the two subject-specific frontiers.

Figures 2.5.1 and 2.5.2 show four plausible outcomes from the MRD described by Wong, Steiner, and Cook (2010). The response surfaces are defined as:

Model 0: $M_{i,t} = 0.5\,R_{i,t-1} + M_{i,t-1} + \varepsilon_i$

Model 1: $M_{i,t} = 4\,Z_{0i} + 0.5\,R_{i,t-1} + M_{i,t-1} + \varepsilon_i$

Model 2: $M_{i,t} = 4\,Z_{0i} + 0.5\,R_{i,t-1} + M_{i,t-1} - 0.05(Z_{1i} * M_{i,t-1}) + 0.55(Z_{3i} * R_{i,t-1}) - 0.025(Z_{1i} * R_{i,t-1} * M_{i,t-1}) - 0.005(Z_{3i} * R_{i,t-1} * M_{i,t-1}) + \varepsilon_i$

Model 3: $M_{i,t} = 4\,Z_{0i} + 2\,Z_{1i} + 2\,Z_{3i} + 0.5\,R_{i,t-1} + M_{i,t-1} - 0.05(Z_{1i} * M_{i,t-1}) + 0.55(Z_{3i} * R_{i,t-1}) - 0.025(Z_{1i} * R_{i,t-1} * M_{i,t-1}) - 0.005(Z_{3i} * R_{i,t-1} * M_{i,t-1}) + \varepsilon_i$,

where *M* and *R* denote normally distributed reading and math scores with pre-test means of zero, standard deviations of 1, and a correlation of 0.2; the assignment and outcome scores are indexed to time with *t*-1 representing the time at pre-test measurement at *t* representing time at post-test; *Z* denotes treatment; and

$Z_1 = 1$ if $M < 0$ and $R \geq 0$ (0 otherwise), $Z_2 = 1$ if $M < 0$ and $R < 0$ (0 otherwise), and

$Z_3 = 1$ if $M \geq 0$ and $R < 0$ (0 otherwise). Model 0 (top left of figure) is not described by the authors, but it is important to consider because it represents either an ineffective intervention or multiple pre-test observations. Model 1 (top right) represents a constant LATE. Model 2 (bottom left) represents a heterogeneous LATE that varies continuously along the union of subject-specific treatment frontiers. And Model 3 represents a heterogeneous and discontinuous LATE. It also represents what one might see if three

different treatments had been administered [depending on which cut score(s) determined assignment] instead of one treatment (regardless of which cut score determined assignment). The perspective plots show a side angle of the true response surface but no observations; the contour scatterplots show the same surface from above so that sampled observations can also be shown.

Figure 2.5.1. Perspective plots of true multivariate RD surfaces

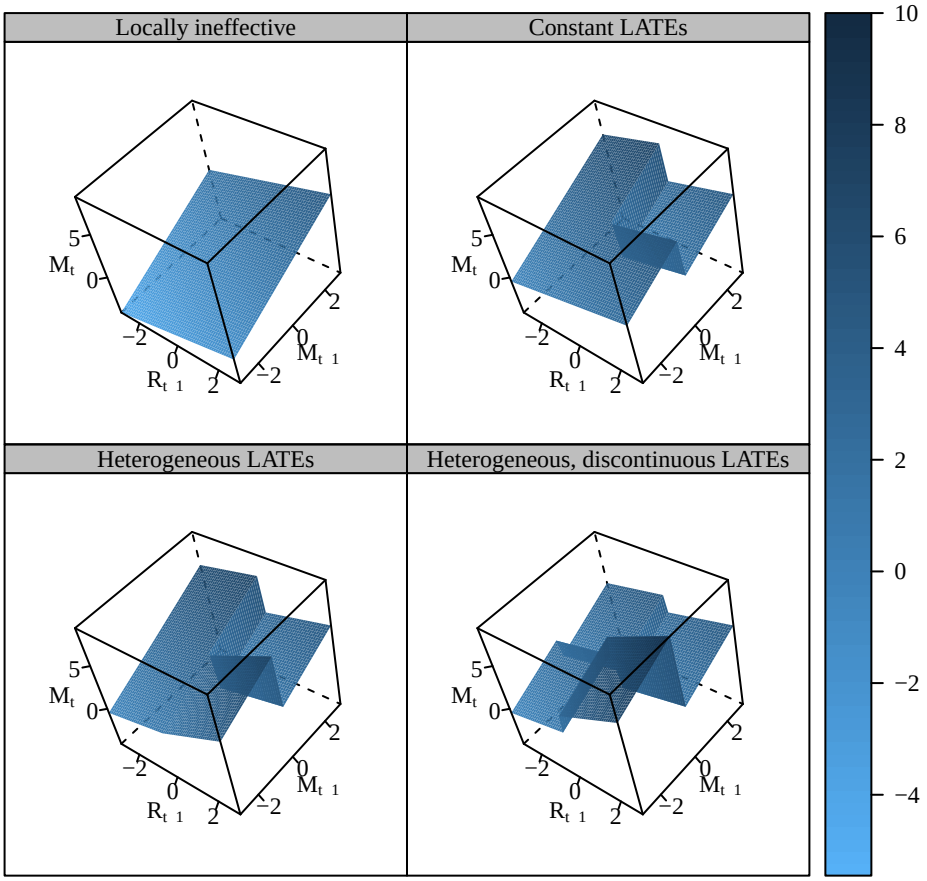
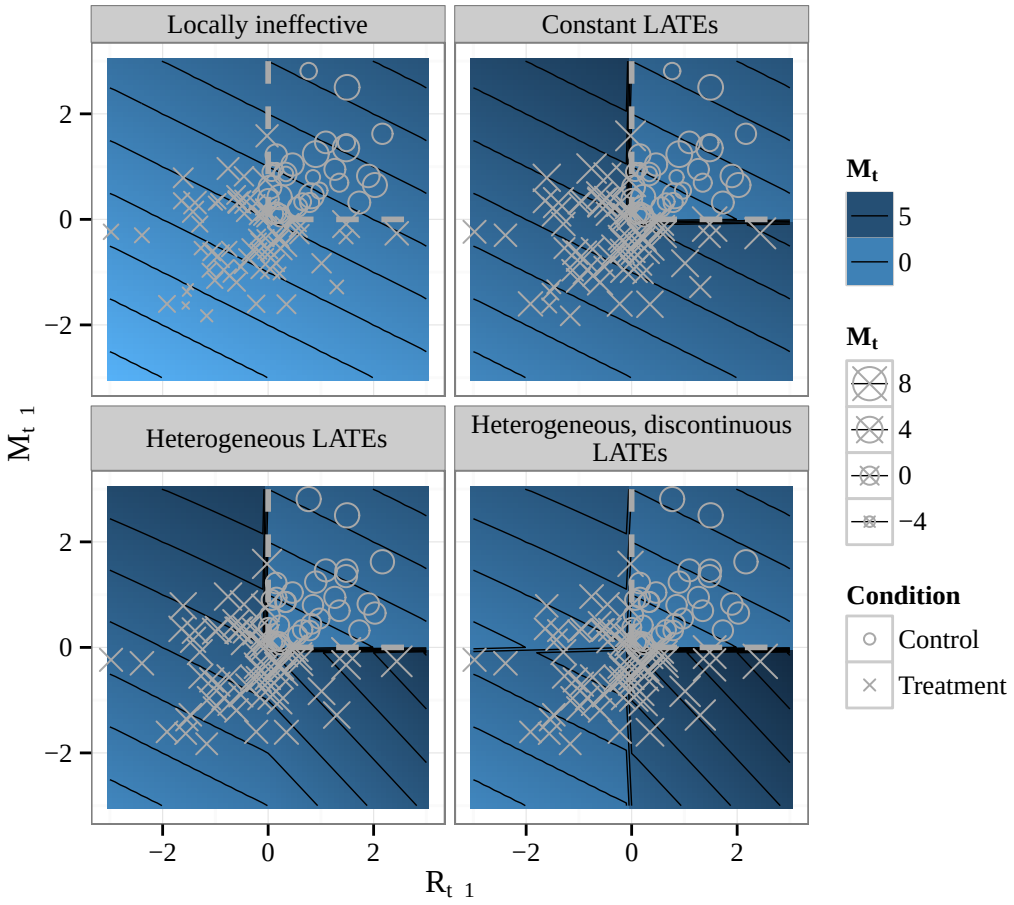


Figure 2.5.2. Contour scatterplots of true multivariate RD surfaces and sampled observations



Wong, Steiner, and Cook (2010) go on to describe four approaches to estimating a weighted average of LATEs along subject-specific frontiers (i.e., restricted cutoffs), as well as results of a Monte Carlo study comparing the approaches. The *frontier approach* nonparametrically estimates LATEs along both frontiers simultaneously and uses weights to estimate an overall LATE. The *centering approach* estimates an overall LATE (not frontier-specific LATEs) by regressing each unit's outcome only on the assignment variable with the cut score closest to that unit [i.e., by ignoring the assignment variable

with the more distal cut score(s)], effectively reducing multiple assignments to a unidimensional mechanism. The *univariate approach* estimates each frontier-specific LATE separately, followed by a weighted overall LATE. The *instrumental variable approach* leverages theory about noncompliance with RD cutoffs and corresponding estimation methods in order to combine estimates of LATEs along the entire cutoff range of each assignment variable (i.e., subsets of the frontiers). Specifically, one assignment variable is used as an instrument for treatment receipt and those assigned by the other variable are treated as misallocated. From the Monte Carlo study, the authors concluded that "given correct model specifications, all four approaches estimate treatment effects without bias, but the instrumental variable approach has severe limitations in terms of more stringent required assumptions and reduced efficiency" (p. 2).

Since the correct model specification cannot be known in practice, Papay, Willet, and Murnane (2011) suggest starting with a full model. The model they recommend qualifies as the univariate approach described by Wong, Steiner, and Cook (2010). Using the latter authors' estimate and notation, the initial model specification recommended by the former authors would yield estimates of four surfaces (one per quadrant) and four heterogeneous and discontinuous LATEs:

$$M_{i,t}=\beta_0+\beta_1 Z_{1i}+\beta_2 Z_{3i}+\beta_3 Z_{2i}+\beta_4 R_{i,t-1}+\beta_5 M_{i,t-1}+\beta_6(R_{i,t-1}*M_{i,t-1})$$
$$+\beta_7(R_{i,t-1}*Z_{1i})+\beta_8(M_{i,t-1}*Z_{3i})+\beta_9(R_{i,t-1}*Z_{3i})+\beta_{10}(M_{i,t-1}*Z_{1i})$$
$$+\beta_{11}(R_{i,t-1}*M_{i,t-1}*Z_{1i})+\beta_{12}(R_{i,t-1}*M_{i,t-1}*Z_{3i})+\beta_{13}(R_{i,t-1}*Z_{2i})$$
$$+\beta_{14}(M_{i,t-1}*Z_{2i})+\beta_{15}(R_{i,t-1}*M_{i,t-1}*Z_{2i})+\varepsilon_i \quad .$$

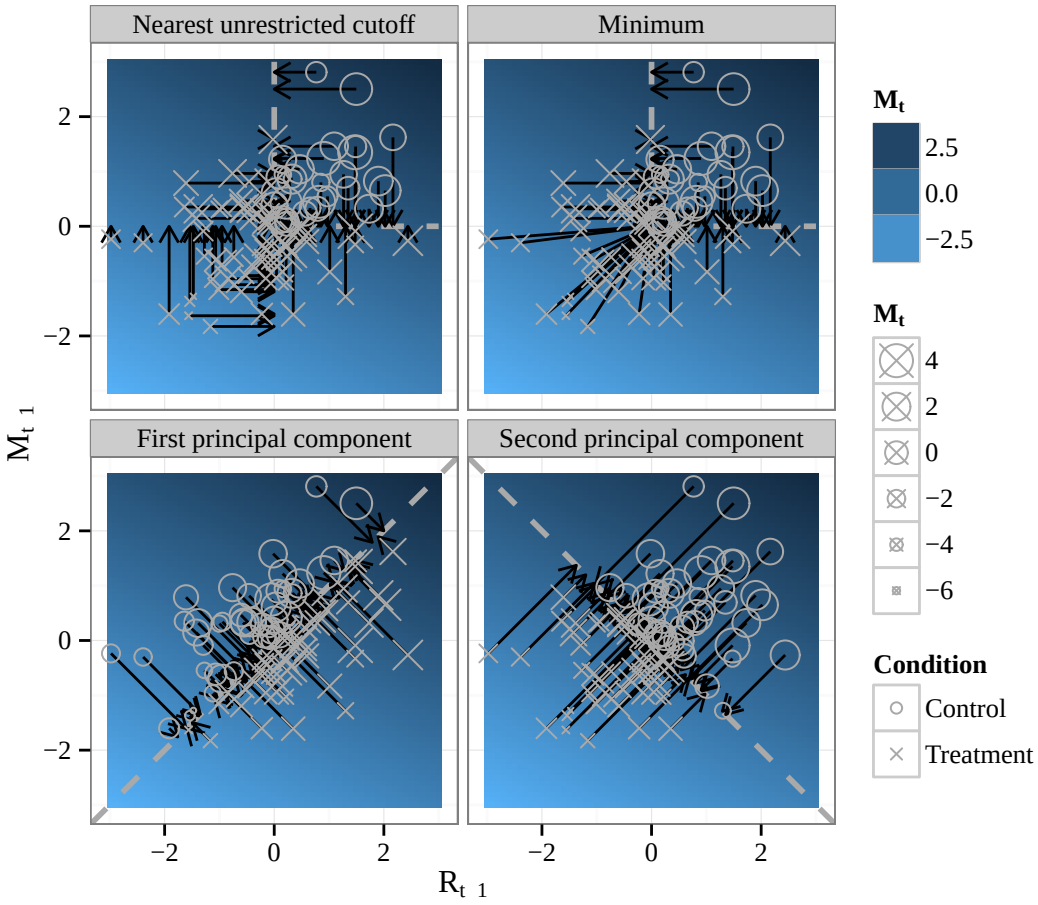
A weighted average of the overall LATE could be calculated from the frontier-specific LATEs. Papay, Willet, and Murnane (2011) go on to describe how to find an optimal

bandwidth h_s^* for local linear regression, although they recommend using ordinary least squares estimation with a subsample of observations exhibiting assignment values less than or equal to h_s^* from the cut score. Similarly, Shadish et al. (2002) recommend considering truncating highly distal observations in order to simplify model specification. They also recommend erring on the side of specifying a polynomial functional form, followed by a backward elimination strategy, in order to avoid under-specification. It bears repeating here that RD designs are not as statistically powerful as randomized experiments. Estimating Papay, Willet, and Murnane's (2011) proposed model with a polynomial specification and excluding distal observations would require a far larger sample size than a randomized experiment. Adding multiple pre-tests to the design would allow one to explore the appropriateness of a parsimonious initial model specification, such as a model that only estimates a constant overall LATE along a simple linear surface.

Cook (2010) describes the benefits of pre-test observations when implementing a RD study. The same advantages hold for MRD. There are several different options for prospectively designing multivariate regression discontinuity studies and measuring distances from cutoffs. Figure 2.6 shows four different options. The options shown in the plot are especially suited to estimating an overall LATE via the centering (minimum distance) approach. All of the response surfaces in the plot correspond to model 0 specified above and represent pre-test observations of the outcome of interest. The dashed grey lines indicate the cutoffs, and the solid black lines show minimum distances from randomly sampled observations to the cutoff. The first design (top left) is described

extensively by Wong, Steiner, and Cook (2010) and Papay, Willet, and Murnane (2011). The other three designs, not discussed by the authors, show that alternative ways of measuring distance to and setting cutoffs for the purpose of treatment assignment. The nearest unrestricted cutoff approach adheres to a design in which a frontier is defined by the union of two (or more) univariate cutoff scores, and distance measured to the nearest cutoff score regardless of its location on the frontier. The minimum distance approach adheres to the same design as the nearest unrestricted cutoff, but for participants who fall below both cutoffs, distance is measured to the nearest location on the frontier instead of the unrestricted cutoff. The minimum distance calculation reflects the fact that a participant would have to gain both reading and math skills (i.e., they have further to go) in order to reach the frontier. The principal component approaches could only be implemented by prospectively designing a RD study that reduces the assignment dimensions. Distance to first principal axis (i.e., first principal component score) provides the most statistical power and closely mirrors random assignment by including a larger variety of persons in both conditions, which may strengthen external validity beyond the study. Distance to the second principal axis (i.e., second principal component score) uses the combination of assignment variables that explain the most variation and allows persons with less of the latent trait to be assigned to the treatment, perhaps increasing rapport with program administrators and participants and preventing fuzzy RD.

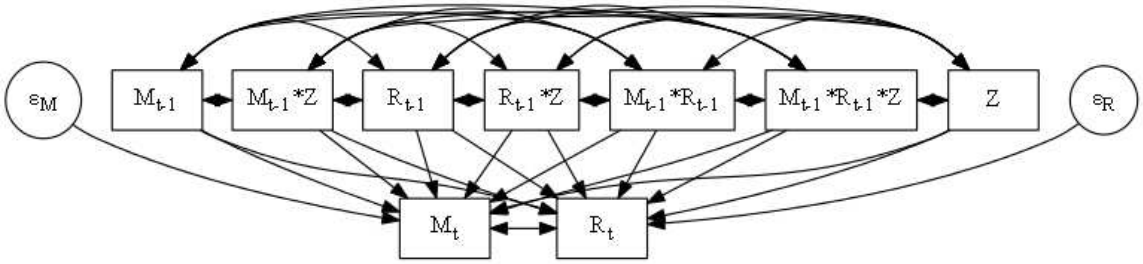
Figure 2.6. Contour scatterplots: Distances from prospectively designed multivariate RD studies



Wong, Steiner, and Cook's (2010) use of the term "multivariate" refers to multiple assignment variables and not necessarily to multivariate analysis of multiple dependent variables. Fully multivariate regression discontinuity analyses are possible and could be highly useful for establishing validity if the treatment is meant to cause multiple outcomes or if the treatment could have caused another outcome of interest (i.e., a non-equivalent dependent variable). Figure 2.7 illustrates the centering approach to a fully

multivariate regression discontinuity path analysis for the math and reading assignment design discussed above.

Figure 2.7. Path diagram of a fully multivariate regression discontinuity design



Spatial regression discontinuity design

Regression discontinuity analysis can be extended to cases of treatment assignment based on geographic borders. Participants in program evaluation studies commonly receive new program services or experience policy changes because they reside in a particular city, school district, or state and not because they were randomly assigned. From a regression discontinuity perspective, the program or policy change represents the treatment, the border surrounding a treatment area represents the deterministic cutoff, and subjects' distances from the nearest border represent the continuous assignment variable. The step of measuring subjects' distances to the nearest treatment border effectively reduces the number of assignment dimensions and qualifies as the centering approach described by Wong, Steiner, and Cook (2010). If the treatment is implemented in multiple geographic areas, then reducing spatial coordinates to minimum distance also effectively reduces unconnected borders to a single border cutoff.

Like MRD, spatial regression discontinuity (SRD) design assumes outcomes vary over a multidimensional surface referenced to multiple treatment assignment variables,

but SRD design differs importantly from MRD. SRD design involves estimating geographically local average treatment effects (GLATEs) at borders of fully enclosed areas. From a MRD perspective, one might say that a SRD design involves estimating LATEs at multiple cut scores along multiple assignment variables, as illustrated in Figures 2.8.1 and 2.8.2. [The figures mirror Figure 2.5.1 and 2.5.2, with models 1-3 coming from Wong, Steiner, and Cook's (2010). The key difference is that Figure 2.8.1 and 2.8.2 illustrate a design with two cut scores per assignment variable instead of one.] However, MRD makes a strong assumption that only one of the assignment variables at a time can vary along the cutoff frontier (Wong, Steiner, & Cook, 2010). SRD accommodates natural experiments by allowing the assignment variables (i.e., the geographically referenced spatial coordinates) to co-vary irregularly in adherence to geographic borders. Irregular borders preclude using the univariate and instrumental variable estimation approaches. Irregular borders could be avoided by prospectively designing a SRD study with new borders that adhere to a straight line for the purpose of the study. Where MRD and univariate RD assume that the assignment variables possess a natural order, SRD assumes the assignment variables (e.g., geographic coordinates) are inherently unordered (Diggle, 2004); order is imposed by calculating distance to the border cutoff. Lastly, MRD does not assume spatial autocorrelation, whereas SRD recognizes and accounts for some degree of spatial autocorrelation by allowing outcomes to vary with distance. Further accounting of spatial autocorrelation can be accomplished with residual covariance structures or spatial lags.

Figure 2.8.1. Perspective plots of true RD surfaces: Prospective design with multiple assignment variables and multiple cutoffs per assignment variable

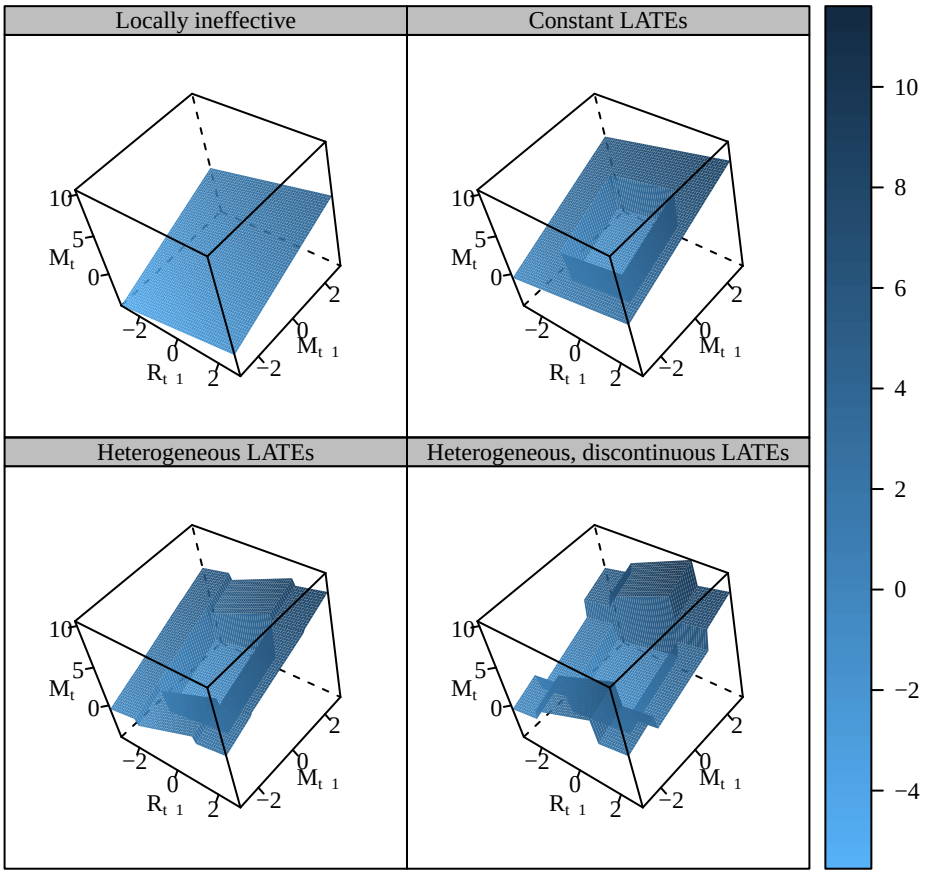
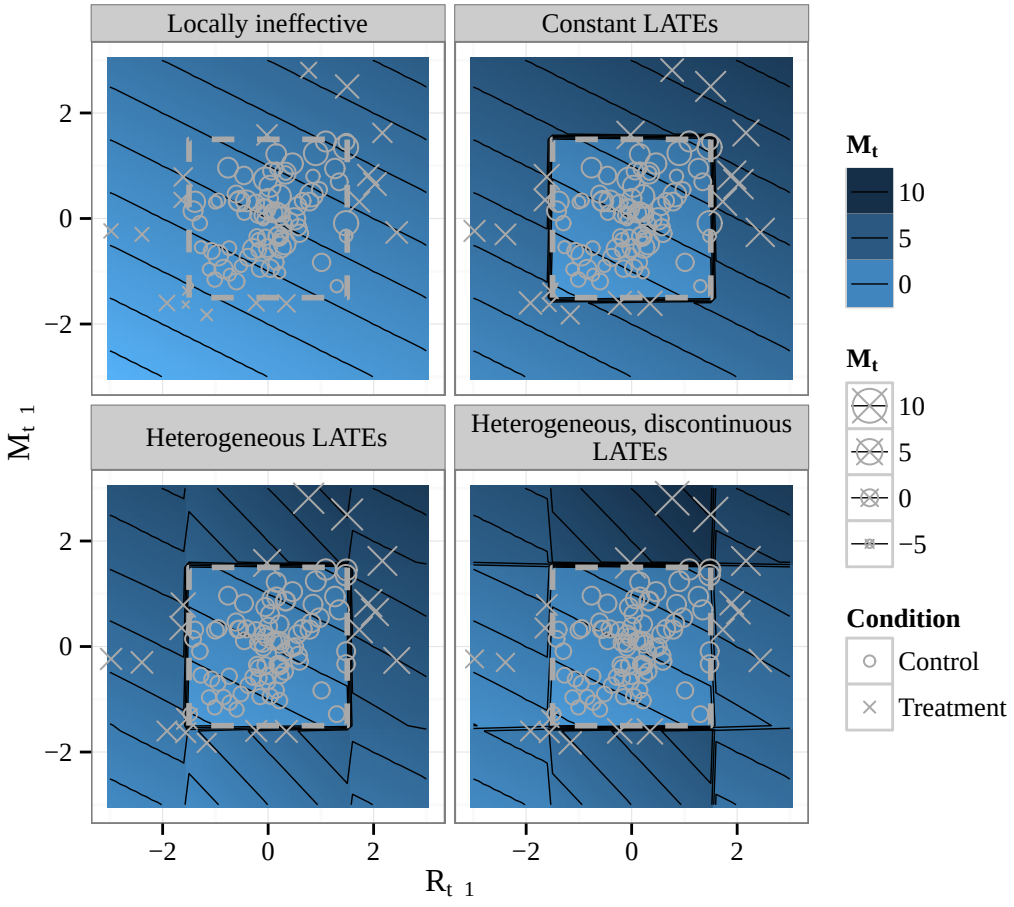


Figure 2.8.2. Contour scatterplots of true RD surfaces and sampled observations: Prospective design with multiple assignment variables and multiple cutoffs per assignment variable



Examples of spatial regression discontinuity designs

A search of extant literature revealed few spatial regression discontinuity (SRD) studies, but the rate appears to be increasing. Most SRD studies have estimated economic effects, and none have examined the educational effects of a specific spatially/geographically implemented program or policy.

Holmes (1998) examined the influence of labor union policies on manufacturing activity at borders separating "probusiness" and "antibusiness" states. Holmes chose a

spatial regression discontinuity analysis in order to "distinguish the effects of state policies from the effects of state characteristics that have nothing to do with state policies" (p. 670), such as agricultural and transportation revolutions, union avoidance, and the advent of air conditioning. A traditional approach would have required a difficult task of statistically controlling for natural resource variation and agglomeration economies. At state borders, natural resources and agglomeration benefits are approximately the same on both sides. County-level manufacturing outcomes were regressed on each county's minimum distance (linear specification) from a pro-/anti-business border, a dummy variable representing pro-/anti-business policies in the county's state, and variables representing each count's projected location *along* the nearest border. Holmes found significantly large and abrupt differences in manufacturing at the state borders. The differences decreased to zero at a distance of a hundred miles from the border and were robust to different specifications. Holmes regressed a theoretically nonequivalent dependent variable—the productivity of industries outside manufacturing —on the same variables explaining manufacturing productivity and found insignificant differences at state borders, thereby strengthening the study's internal validity. Holmes cautions that even though border discontinuities are evident in a manner consistent with prior theory, the local treatment effects may be attributable to geophysical discontinuities, prior state policies, or policies other than current labor union policies (i.e., history).

Black (1999) examined the localized effect of test scores on home values at school districts borders in order to estimate the monetary value of school improvement while controlling for the influence of neighborhood characteristics that confound and typically overestimate the value of better schools (i.e., because better schools tend to be

located in more affluent neighborhoods). Black's model does not formally include distance as a treatment assignment variable. Instead, the sample was restricted to houses within 0.35 mile from the nearest border, and housing prices were regressed on test scores, control variables (e.g., number of bathrooms), and dummy variables for each border segment shared by two school districts. The localized estimates of the effect of test scores on housing prices were significantly positive but smaller in magnitude compared to the full sample without boundary dummy variables, as Black theorized. Test scores did not significantly explain measures of home size, qualifying them as nonequivalent dependent variables and strengthening the study's internal validity.

Bayer, Ferreira, and McMillan (2007) replicated Black's (1999) study in a different area and with richer data, but with distance measured with respect to centroids (i.e., centers of geographic areas). Like the latter author, they found that adding demographic control variables and limiting inferences to geographic borders, where the quality of housing is assumed similar on both sides of a border, reduced the estimate of school quality's influence on housing prices. The reduction occurred within their study and relative to studies with weaker designs. The authors conducted thorough exploratory analyses of border discontinuities to assess the degree of selection with respect to the border, which they call "sorting". Geographic sorting in this context may be thought of as residential arbitrage, arising from knowledge of a pre-existing cutoff. They found clear discontinuities in housing prices (their outcome of interest) and in demographics, but not in housing quality. Those findings stress the lack of control a researcher has over SRD natural experiments with known borders. They also reinforce the key SRD theory that geographically local estimation can rule out competing explanations for the observed

treatment effect by limiting inferences to where participants are similar except for the treatment (i.e., good counterfactuals at the border). In order to avoid omitted variable bias, the authors statistically controlled for demographics when regressing housing prices on the school quality treatment variable determined by the border.

Moore (2009) used SRD to re-analyze data from Card and Krueger's (1994) study of fast food restaurants in Pennsylvania and New Jersey before and after New Jersey raised its minimum wage. Economic theory asserts that increasing minimum wage should cause employment to decrease, but they did not find a (non-local) average treatment effect on employment resulting from minimum wage increase. Like Holmes' (1998) use of SRD, the goal of the re-analysis was to estimate the GLATE of a new policy representing an exogenous treatment. This contrasts with Black (1999) and Bayer et al.'s (2007) use of SRD to describe/quantify behavior with no treatment manipulation. Because policy changes have an intended outcome, Card and Krueger (1994), Holmes (1998), and Moore (2009) were able to add non-equivalent dependent variable design and analytic enhancements to help rule out history and other validity threats. The employment outcome in Card and Krueger's (1994) study represented the unintended, non-equivalent dependent variable, but it was the most pertinent to the research question regarding minimum wages influence on employment. Pre-test exploratory analyses revealed a discontinuity in the wage outcome before the policy change, requiring a difference-in-differences analysis. The pre-test analysis also revealed substantial design effects due to intraclass correlation/spatial autocorrelation within geographic subregions, requiring multilevel (hierarchical linear) modeling to obtain design-based standard errors. As expected, raising the minimum wage (i.e., being located in New Jersey) had a

significant, positive effect on starting hourly wages. The wage outcome was not significantly correlated with minimum distance from the border, which is to say the estimate of the average treatment effect was not local (see Figure 2.9). There was not enough evidence to conclude that raising the minimum wage had a local average treatment effect on the number of full-time equivalent employees at the state border. The finding of a statistically insignificant GLATE was consistent with Card and Krueger's (1994) finding with respect to the non-local average treatment effect. Moore (2009) noted that the findings could be sensitive to how distance to the border is defined and showed that distance can be defined differently depending on the subregion within which a location is clustered. That is, distance can be rescaled within subregions to reflect socioeconomic distances. For example, the correlation between latitude and longitude of locations within a subregion could be used to find other axes and stretch/shrink them in order to calculate Mahalanobis distances (see Figure 2.10).

Figure 2.9. Fitted line plots illustrating inferences from a SRD analysis by Moore (2009)

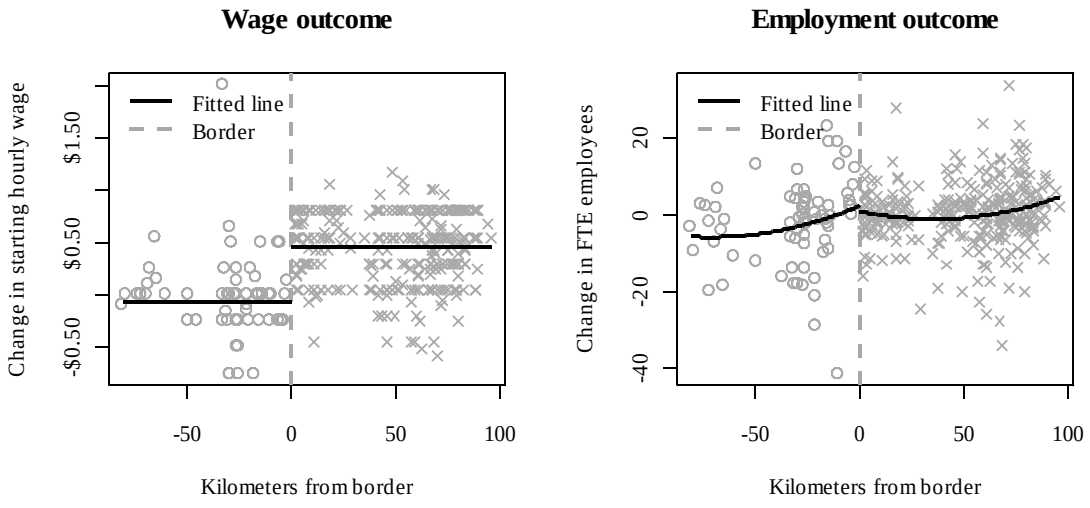
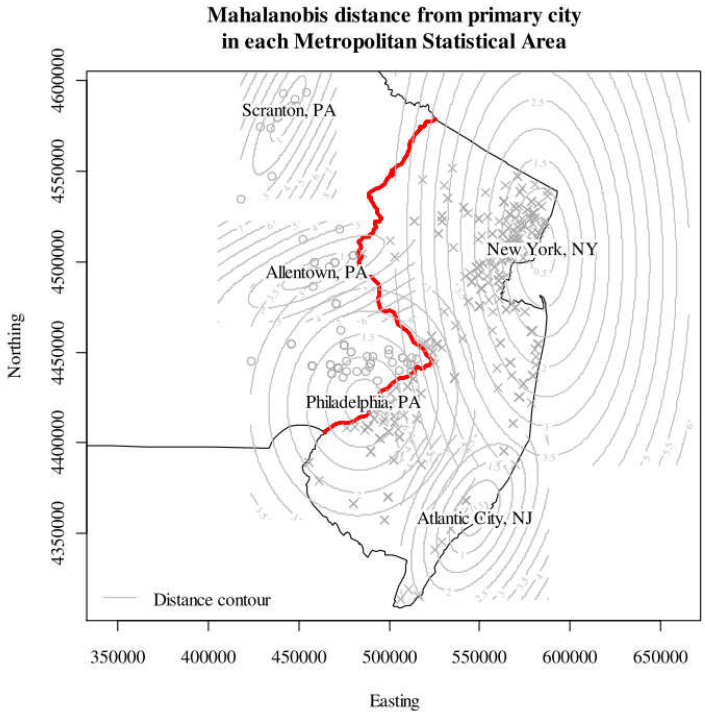


Figure 2.10. Subregion distances operationalized as Mahalanobis distances



Keele and Titiunik (2011) describe many of the strengths and weaknesses of spatial regression discontinuity designs relative to simple RDs and propensity score matching. They do so in the context of estimating the effect of ballot initiatives' on voter turnout. In particular, they describe how SRD natural experiments are more analytically demanding than simple RDs because spatial regression discontinuity designs identify an infinite number of GLATEs and known borders allow selection. The authors claim that GLATEs must be defined as curves instead of a single parameter. This is analogous to Wong, Steiner, and Cook's (2010) claim that multivariate regression discontinuity design requires estimating LATEs along frontiers but without strongly assuming straight-line frontiers. They point out that reducing two-dimensional space to one-dimensional distance will effectively place two points close together when they are actually very far

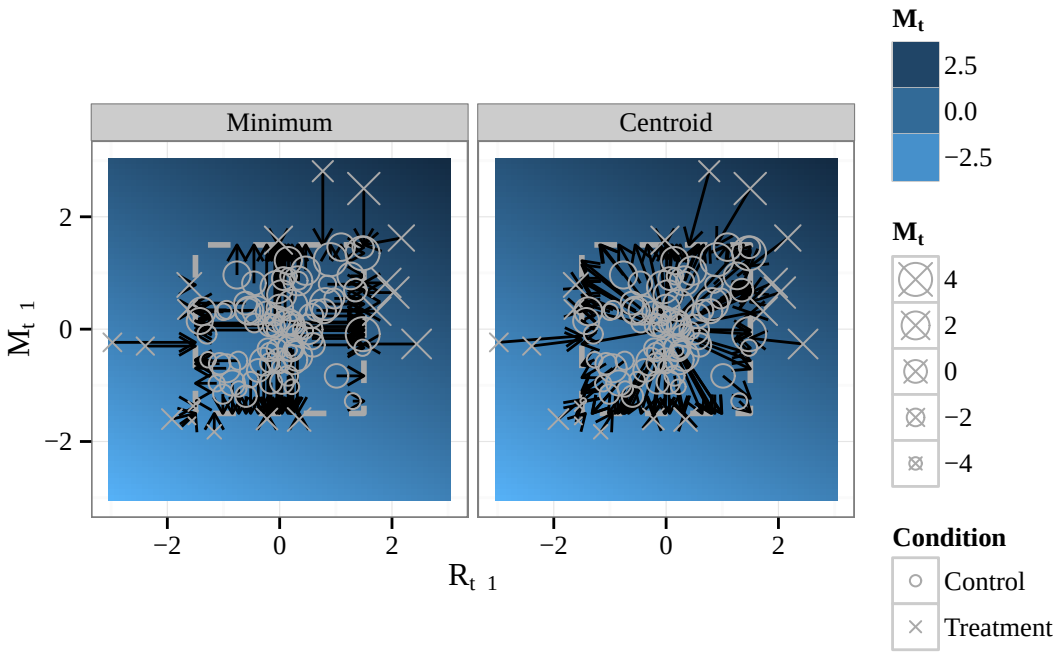
apart spatially/geographically. It is unclear why that would be problematic as long as participants on both sides of the border are good counterfactuals *on average*, as described by Holmes (1998). Assessing and addressing selection requires the researcher to consider the local selection mechanism and to exclude borders and cases that violate RD assumptions. In other words, SRD designs may not be any stronger than covariate adjustment designs (e.g., propensity score matching), and internally valid inferences may not generalize to other localities. They propose a formal strategy for testing the hypothesis that propensity score matching is a stronger design than SRD, the former of which is widely considered to be weaker and more analytically demanding than simple RD (Shadish et al., 2002). It is unclear why they do not consider the possibility of simply including covariates in the SRD model, as Black (1999), Bayer et al. (2007), and Moore (2009) did, and especially in light of findings by Shadish, Clark, and Steiner (2008) that the inclusion of control variables in a regression model can effectively reduce bias as much as propensity score matching. Nevertheless, by framing SRD and propensity score matching as distinct choices and applying an inference strategy to voter turnout data, Keele and Titiunik (2011) find that accounting for distance effectively leads to covariate balance, and they conclude that SRD is preferable to matching in their example. In their analysis, they exclude borders and corresponding cases that appeared to violate assumptions of pre-test continuity.

The above SRD studies reveal some steps that might be worthy of emulation and some steps that could be improved in future spatial regression discontinuity studies. All the studies measured distance to the border, but distance was operationalized in a couple different ways. In all but one study, distance was measured directly to the nearest point

along the border (i.e., the frontier), similar to taking progressively larger buffers around the treatment area until the buffer reaches a distance considered overly inclusive of distal observations. In one case (Bayer, Ferreira, & McMillan, 2007), distance to the border was measured either directly toward the centroid of the treatment area if in the control group or away from the centroid if in the treatment group. The centroid approach does not yield the minimum possible distance, but it may improve covariate balance to the degree that variation in confounding variables radiate from the center of an area. The two ways of operationalizing distance are shown in Figure 2.11. All of the studies involved some form of restricting the sample to a smaller range around the border. All of the studies conducted extensive validation analyses above and beyond estimation of GLATEs, such as pre-test explorations of border continuities, analyses of sensitivity to model specifications, and modeling nonequivalent dependent variables. Some of the studies excluded borders (and corresponding cases) that were common to more than one area (e.g., borders that double as a school district *and* city border) in order to mitigate the selection and/or history validity threats. Some of the studies treated distance as an ordinal measure and performed multiple statistical tests rather than simply using regression analysis to estimate and discuss the magnitude of the effect as a continuous and possibly curvilinear function of distance form the border. None of the studies attempted to generalize GLATEs to areas beyond the borders or to the univariate or instrumental variable estimation approaches because a geographic border is rarely meaningfully constant across the range of other dimensions (i.e., latitude and longitude coordinates vary in irregular patterns). Given that Wong, Steiner, and Cook's (2010) simulation study showed that the centering estimation approach yielded similar estimates

to other estimation approaches, then using a measure of minimum distance to reduce multidimensional space to a unidimensional assignment mechanism has clear advantages. Even though the general aim of RD studies is to estimate LATEs, one may wish to estimate or predict heterogeneous GLATEs at specific locations along the border, in which case the models proposed by Holmes (1998) and Keele and Titiunik (2011) offer guidance.

Figure 2.11. Contour scatterplots of true RD surface and sampled observations: Operational determinants of direction and distance



Discussion

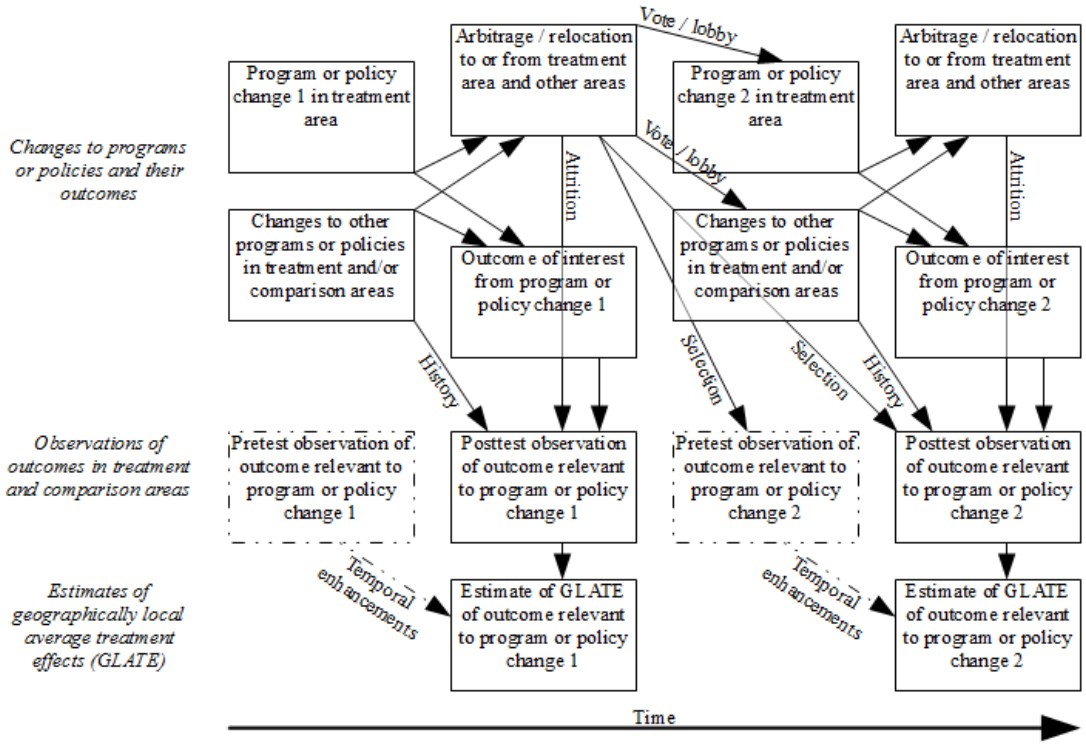
Like basic RD, the strengths of SRD include a high degree of feasibility and a high degree of knowledge about the treatment assignment mechanism. SRD is widely applicable due to the widespread practice of implementing programs and policies in

geographic areas, such as school districts or states, and the likelihood of geographic variation of intended outcomes before and after implementation. For example, if one state implements teacher performance pay and a neighboring state does not, then a student's location relative to the state border *determines* their exposure to the treatment. In theory, SRD is just another version of the basic RD in which the outcome of interest covaries with distance, a border sharply determines who receives the treatment condition, and participants near each other on opposite sides of the border are highly similar (i.e., ignorably different) in all regards except for the treatment. If the basic assumptions hold, then SRD designs will not offer as much simplicity, statistical power or generalizability as randomized experiments, but estimates of overall local average treatment effects will be unbiased. Additionally, SRD designs will offer clear advantages over quasi-experimental designs in which the treatment assignment mechanism is unknown and must be modeled and/or designs that lack a comparison group. Given that Wong, Steiner, and Cook's (2010) Monte Carlo study showed that the centering approach can yield unbiased frontier average treatment effects and that several SRD studies have found that individuals near each other on opposite sides of the borders do tend to be similar (i.e., qualify as counterfactuals), then it seems reasonable to conclude that SRD designs, although not as simple, possess some of the same strengths of basic RDs. If a SRD study was designed prospectively with new borders, then it would possess nearly all of the same strengths as a basic RD.

Even though SRD designs are widely applicable and possess some of the same strengths of basic RD designs, SRDs face several validity threats. As shown in Figure 2.12, selection, history, and attrition threaten the internal validity of GLATE estimates

when preexisting borders determine assignment to the treatment condition. If an established, well-known border is used for convenience, then self-selection in the form of geographic arbitrage (i.e., re-locating to maximize economic gain and minimize costs) could invalidate causal conclusions. Given that economic arbitrage/sorting at borders was found in at least two SRD studies (Black, 1999; Bayer, Ferreira, & McMillan, 2007), then selection seems highly plausible when considering the effect of an intervention on some outcome other than arbitrage (i.e., the outcome that the intervention or policy is meant to affect). History can also threaten the validity of GLATE estimates because administrators and policy makers may geographically implement several programs simultaneously to affect a common outcome (e.g., student achievement). Additionally, attrition can threaten the validity of GLATE estimates when a program or policy change causes relocation to a different area, but high costs of moving may deter attrition. Lastly, all SRD studies make strong assumptions about minimum distance. As Keele and Titiunik (2011) discuss, there is actually an infinite number of distances along the treatment border and an infinite number of GLATEs. Choosing how to operationalize and measure distance to the border is not unlike the RD task of choosing a functional form or bandwidth: the best choice is unknown, but the validity of inferences depends on making a good choice.

Figure 2.12. Overview of the process by which preexisting borders allow threats to the internal validity of GLATE estimates over time



Weaknesses of SRD studies can be avoided via design enhancements or accounted for via analysis. Design-based enhancements could help rule out selection and history threats and provide greater statistical power. An ideal SRD design would create and enforce a new geographic border cutoff just for the program or policy of interest, which would provide full knowledge of the assignment process and reduce the confounding influence of other programs. A researcher could perform a principal components analysis of geographic coordinates and set the cutoff at the first principal axis in order to maximize statistical power and variety of persons in both conditions, with each location's principal component score representing centered distance from the cutoff. However, excluding some people from a potentially beneficial service or policy because of where

they live could provoke public backlash. For this reason, a researcher who establishes new boundaries for an SRD design might be wise to consider providing the treatment to persons in the control group after the study has run a sufficient course for estimating the LATE. Analytic enhancements are less preferable to design-based enhancements, but they can help adjust for selection. All SRD studies should be actively validated by conducting pre-test analyses of selection and post-test analyses of misallocation. If pre-test data were collected by design, then an evaluator could determine the degree to which a GLATE estimate at a preexisting border represents an improvement over a non-local estimate. The first step would involve estimating GLATEs for the pre-test outcome and variables representing socioeconomic conditions. If distance and either the pre-test or socioeconomic GLATE are significantly different than zero, then estimating a post-test GLATE would be more appropriate than estimating a non-local effect. Pre-test data could also be used to select an appropriate measure of distance to the border. The best measure of distance is perhaps the one that results in the highest degree of pre-test comparability at the border. If the pre-test GLATE is significantly different than zero, then that would indicate arbitrage and warrant estimating a difference-in-differences GLATE to adjust for selection (Cook, 2008). According to Shadish et al. (2002), pre-test data can help strengthen statistical conclusion validity by empirically guiding the choice of an appropriate functional form for the relationship between the assignment and outcome variables. If multiple pre-test and post-test data were collected by design, then one could estimate change in GLATEs over time. Here, the first GLATE resembles an interrupted time series, and the counterfactual line is informed by (i.e., carries forward) the pre-test trend. Under RD, it is not necessary to include control variables in the

piecewise regression model because the cutoff completely determines assignment. When estimating a GLATE at a preexisting border, however, controlling for variables that explain selection or variation in socioeconomic conditions is warranted and may reduce bias. Table 2.2 summarizes SRD validity threats and corresponding enhancements.

Table 2.2. Summary of SRD validity threats and corresponding enhancements

Validity threat	Design and analytic enhancements that may help rule out validity threat	Sources describing threat and/or enhancement
Fuzzy treatment assignment (i.e., misallocation, selection, sorting, manipulation, attrition) violates the RD design and threatens internal validity and statistical conclusion validity by biasing LATE estimates.	• Prospectively design and control RD studies by establishing and enforcing a new cutoff.	• Cook (2010) • Lee and Lemieux (2010) • Moore (2009)
	• Use pre-test data to assess ignorability by empirically examining continuity of the outcome variable and covariates at the border. • Assess ignorability at post-test by calculating the percent of cases that were misallocated.	• Bayer et al. (2007) • McCrary (2008) • Moore (2009) • Keele and Titiunik (2011) • Urquiola and Verhoogen (2009)
	• Conduct a difference-in-differences analysis if pre-test discontinuities in the outcome are evident at the border. • Add covariates to analytically control for selection if nonignorable.	• Bayer et al. (2007) • Black (1999) • Lee and Lemieux (2010) • Moore (2009) • Shadish et al. (2008)
	• Use an instrumental variable approach or settle for estimating an intent-to-treat LATE.	• Shadish et al. (2002) • Imbens and Lemieux (2007)
An incorrectly specified functional form threatens internal validity and statistical conclusion validity by biasing (G)LATE estimates.	• Conduct a specification search using pre-test observations. • Conduct sensitivity analyses of chosen functional form.	• Moore (2009) • Shadish et al. (2002)
	• Over-specify the model, followed by backward elimination if study is statistically powerful.	• Papay, Willet, and Murnane (2011) • Shadish et al. (2002)
	• Use nonparametric estimation (i.e., do not specify a functional form).	• Hahn et al. (2001)
	• If using nonparametric estimation, empirically choose a bandwidth that cross-validates.	• Imbens and Lemieux (2007)

Validity threat	Design and analytic enhancements that may help rule out validity threat	Sources describing threat and/or enhancement
	• Exclude distal observations to simplify specification.	• Bayer et al. (2007) • Black (1999) • Shadish et al. (2002)
History threatens internal validity and statistical conclusion validity by confounding two or more treatments, thereby biasing (G)LATE estimates.	• Prospectively establish a new cutoff that will not co-determine assignments to other treatment conditions.	• Shadish et al. (2002)
	• Exclude segments of geographic boundaries that are shared by more than one governmental or relevant service entity.	• Bayer et al. (2007) • Black (1999) • Keele and Titiunik (2011)
Local estimation (relying on observations near the cutoff) and collinearity (between assignment and treatment variables) threaten statistical conclusion validity by lowering statistical power.	• Set cutoff at or near mean of assignment variable	• Cappelleri et al. (1994) • Goldberger (1972) • Shadish et al. (2002)
	• Design a longitudinal SRD study that collects abundant data from multiple time periods.	
RD designs limit internally valid estimates to the cutoff, thereby limiting external validity of generalizations both within and beyond the study, especially if estimates are geographically local.	• Estimate pre-test fitted lines and extrapolate beyond the cutoff if warranted.	• Campbell (1969) • Keele and Titiunik (2011) • Moore (2009) • Shadish et al. (2002)
	• Estimate change in (G)LATEs over time [i.e., to see if the initial (G)LATE generalizes to other time periods].	

Q Comp: A spatial regression discontinuity natural experiment

The Q Comp program qualifies as a spatial regression discontinuity design. Q Comp's multiple levels of participation and implementation could be considered a challenge for evaluating the program, but it can be leveraged using SRD. The program has been implemented geographically, with school district borders determining schools' participation in Q Comp and distance acting as a continuous assignment variable. Given that socioeconomic characteristics are spatially auto-correlated (i.e., people reside near those with similar characteristics whether through choice or coercion), it would be reasonable to expect that students and schools near each other on opposite sides of borders are similar. At the district level, it would also be reasonable to expect that districts with higher student achievement would be more inclined to participate in a program like Q Comp that ties pay to student achievement than districts with lower student achievement. However, at the school level teachers in schools with lower student achievement may be less inclined to participate in Q Comp and may have more in common with nearby schools in another district than schools across town with higher student achievement relative to the district average. Even though districts self-select into Q Comp, the border offers a location where student achievement in Q Comp schools can be compared to achievement in non-participating schools via SRD analysis in order to overcome difficulties identified in prior evaluations.

Even though district borders determine participation of schools and local comparability seems plausible, the reality is more complicated. Q Comp qualifies as a natural experiment because it has been implemented using pre-existing geographic borders. The same borders that determine participation have also determined school

resources and many other policies over the years leading up to Q Comp. Additionally, studies by Black (1999) and Bayer et al. (2007) suggest that families are knowledgeable of district borders and practice geographic arbitrage/sorting to get their children into good districts. If a new Q Comp border could have been established to determine participation, then comparability at the border would be highly likely (i.e., ignorably different). However, the use of pre-existing borders threatens the validity of conclusions about Q Comp's effectiveness under the SRD approach. Design and analytic enhancements are necessary to help ensure that observed GLATEs are attributable to Q Comp.

Chapter 3: Methods

Chapter 2 presented an overview of Minnesota's Quality Compensation for Teachers (Q Comp) program and summarized the spatial regression discontinuity (SRD) method. In this chapter, I propose a modeling strategy for estimating Q Comp's impact on student achievement via SRD analysis. A key component of the strategy is to leverage pre-Q Comp data to help rule out validity threats and evaluate the merits of the SRD approach more broadly.

Evaluation questions

The purpose of this dissertation is to answer two sets of questions: a set of substantive evaluation questions regarding Q Comp and a set of methodological questions pertaining to the SRD approach as it applies to school district borders.

1. Has Q Comp been effective? To what degree has Q Comp led to student achievement gains as theorized? Which Q Comp districts and schools added significant value and warrant emulation?
2. Is applying SRD worthwhile (given the effort and costs to parsimony and external validity)? To what degree does SRD yield well-matched comparisons (i.e., counterfactuals) at the Q Comp border on average? To what degree does the SRD approach, with design and analytic enhancements, rule out validity threats?

Data

This study analyzes publicly available school attribute data (student achievement/characteristics and finances) from the Minnesota Department of Education (MDE; 2013) and school and district geographic data from the Minnesota Geospatial Information Office (2013). Only Independent, Common, and Special school districts (the

most common types enrolling the most students) are included. Schools in other types of districts and charter schools have been excluded because they tend to enroll more narrowly defined student groups (e.g., students recovering from substance abuse) and because their organizational structures and student attendance patterns are not geographically defined.

The outcome of interest is academic achievement of students. Student achievement is operationalized as school means of student test scores on the math and reading Minnesota Comprehensive Assessments (MCAs) and the Mathematics Test for English Language Learners (MTELL). [Note that students with cognitive disabilities who took the MCA-Modified or Minnesota Test of Academic Skills (MTAS) alternate assessments are not included in this study. Also note that very small student cohorts are not included because MDE does not publish school mean test scores based on less than 10 students (in order to protect students' privacy).] Minnesota developed the MCAs and MTELL to comply with Title I of the Elementary and Secondary Education Act and for state accountability purposes. Within a grade level and subject, they are standardized in three senses of the word.

1. Items are written and tests are constructed to align with the Minnesota K-12 Academic Standards, which articulate what students are expected to learn.
2. Scores are scaled and equated to allow comparisons within (but not between) academic standard-setting windows to criteria (i.e., proficiency cut scores) and from one year to another.

3. The tests are administered to all students in a consistent manner. Standardizing administration helps minimize and hold constant variations in testing conditions as a source of measurement error.

Standardization leads to a high degree of score reliability (greater than 0.9) at proficiency cuts and decreasing toward the tails of grade-level achievement distributions. Measurement error, although minimized at the student level, remains when scores are aggregated. In other words, the school mean outcomes in this study also contain measurement error and qualify as estimates of a latent student achievement factor (Kane & Brennan, 1977). Schools with fewer students and/or larger shares of very low- or high-achieving students will have less reliable means than schools with larger shares of students near the proficiency cut scores.

The MCAs were first administered to third and fifth graders in 1998 and to students in grades 7, 10, and 11 starting in 2004. Students in grades 4, 6, and 8 began taking the MCAs in 2006, the same year that Q Comp began. The MTELL, first administered in 2007 and discontinued after 2010, was developed to measure the same math construct as the MCAs and on the same scale. As such, MTELL mean scores are combined with MCA mean scores via student-weighted averaging in applicable years.

Scores were transformed and limited to grades 3, 5, 7, 10, and 11 to establish a consistent scale for scores spanning two years before Q Comp (i.e., 2004) through 2013. Transforming scores is necessary because the scales of published scores were criterion-referenced. The scales have changed over time with changes to grade- and subject-specific achievement levels (i.e., across standard-setting windows). Consistency across grades, subjects, and years was accomplished by norm-referencing within each grade,

subject and year. That is, the student-level statewide population mean scale score were subtracted from each school's mean, after which the centered school mean was divided by the student-level statewide population standard deviation to arrive at school means of student z-scores. In addition to scale consistency, the student z-score metric facilitates interpreting parameter estimates as standardized effect sizes.

Student data must be referenced to geographic data for SRD analysis. The Minnesota Geospatial Information Office (2013) provides school location data (points) and school district boundary data (polygons). (Note that some school locations were missing from the Minnesota Geospatial Information Office's data, especially in the earlier years. Missing coordinates were imputed from the next available year or geocoded from published addresses if completely missing.) School location coordinates were merged with student data. Distance to the dissolved Q Comp borders will then be calculated within each year. Lastly, data were combined across years to arrive at the final, longitudinal, spatially-referenced data set.

Methods

Given 1) the characteristics of the Q Comp program, 2) how it qualifies as a SRD natural experiment, and 3) the available data, there are several SRD design and analytic enhancements that can be applied to rule out validity threats identified in chapter 2. Pre-test analysis, covariance adjustment, and checks for misallocation will address the fuzzy treatment assignment threat. Pre-test analysis, sensitivity analysis, and an exclusion criterion will address the functional form threat. Longitudinal design and analysis will be applied to address the statistical conclusion and generalizability validity threats. The history validity threat can not be addressed feasibly. Table 3.1 repeats the validity threats

and best practices from Table 2.2, adding enhancements that are applicable to the study of Q Comp.

Table 3.1. Summary of Q Comp evaluation validity threats and SRD enhancements

Validity threat	Design and analytic enhancements that may help rule out validity threat	Applicability to Q Comp evaluation
Fuzzy treatment assignment (i.e., misallocation, selection, sorting, manipulation, attrition) violates the RD design and threatens internal validity and statistical conclusion validity by biasing (G)LATE estimates.	• Prospectively design and control RD studies by establishing and enforcing a new cutoff.	• The enhancement can not be applied. That is, the Q Comp evaluation was not prospectively designed and participation was not controlled.
	• Use pre-test data to assess ignorability by empirically examining continuity of the outcome variable and covariates at the border.	• The enhancement is applicable because pretest data were collected. It is also important for evaluating the SRD approach.
	• Assess ignorability at post-test by calculating the percent of cases that were misallocated.	• The enhancement is applicable and important to apply because some schools within participating Q Comp districts may not have participated.
	• Conduct a difference-in-differences analysis if pre-test discontinuities in the outcome are evident at the border. • Add covariates to analytically control for selection if nonignorable.	• These enhancements will be applied proactively, regardless of pre-test ignorability, because pre-test data and other controls are available to potentially explain additional variation in outcomes.
	• Use an instrumental variable approach or settle for estimating an intent-to-treat LATE.	• The enhancement will not be applied. The intent-to-treat effect is of interest because it is more externally valid (i.e., misallocation is not entirely avoidable in a widely implemented public program and the program's actual effect is of greater interest than its effect under ideal conditions).

Validity threat	Design and analytic enhancements that may help rule out validity threat	Applicability to Q Comp evaluation
An incorrectly specified functional form threatens internal validity and statistical conclusion validity by biasing (G)LATE estimates.	• Conduct a specification search using pre-test observations. • Conduct sensitivity analyses of chosen functional form.	• The enhancements are applicable and important to apply both for obtaining unbiased estimates of Q Comp GLATEs and for evaluating the SRD approach relative to simpler quasi-experiments. Because pre-test data are available, pre-test exploration will be prioritized over sensitivity checks in order to avoid over-fitting of post-test data.
	• Over-specify the model, followed by backward elimination if study is statistically powerful.	• The enhancement is applicable, although results of pre-test data analyses will guide the choice of a functional form and help avoid over-specification.
	• Use nonparametric estimation (i.e., do not specify a functional form). • If using nonparametric estimation, empirically choose a bandwidth that cross-validates.	• The enhancements will not be applied because parametric estimation via multilevel/mixed effects modeling is desirable due to nesting of time-specific observations within schools and districts.
	• Exclude distal observations to simplify specification.	• The enhancement is applicable and important to apply because many non-participating (comparison group) schools are located much further from the Q Comp border than the furthest participating schools (i.e., because participating schools are bound/land-locked).
History threatens internal validity and statistical conclusion validity by confounding two or more treatments, thereby biasing (G)LATE estimates.	• Prospectively establish a new cutoff that will not co-determine assignments to other treatment conditions.	• The enhancement can not be applied because the Q Comp evaluation was not prospectively designed. As such, inferences about the effects of Q Comp could be confounded with other programs implemented district-wide

Validity threat	Design and analytic enhancements that may help rule out validity threat	Applicability to Q Comp evaluation
	• Exclude segments of geographic boundaries that are shared by more than one governmental or relevant service entity.	• The enhancement will not be applied in this study due to data and resource limitations and because applying this enhancement could drastically reduce statistical power considering the degree to which school districts share borders with cities and counties.
Local estimation (relying on observations near the cutoff) and collinearity (between assignment and treatment variables) threaten statistical conclusion validity by lowering statistical power.	• Set cutoff at or near mean of assignment variable	• The enhancement can not applied because the Q Comp evaluation was not prospectively designed.
	• Design a longitudinal SRD study that collects abundant data from multiple time periods.	• The enhancement is applicable because data were collected from multiple years before and after the Q Comp program was implemented. It is important to apply for statistical power and because Q Comp's effects on student achievement could depend on duration of implementation.
RD designs limit internally valid estimates to the cutoff, thereby limiting external validity of generalizations both within and beyond the study, especially if estimates are geographically local.	• Estimate pre-test fitted lines and extrapolate beyond the cutoff if warranted.	• The enhancement will not be applied proactively. Estimates may be non-local if student achievement does not vary geographically with respect to the Q Comp border (i.e., if a simpler quasi-experiment suffices).
	• Estimate change in (G)LATEs over time [i.e., to see if the initial (G)LATE generalizes to other time periods].	• The enhancement is applicable because data were collected from multiple years before and after the Q Comp program was implemented. It is important to apply for statistical power and because Q Comp's effects on student achievement could depend on duration of implementation.

The applicable design and analytic enhancements will be applied in the following order:

1. Check for misallocation and exclude distal observations from the comparison group of schools not participating in Q Comp.
2. Conduct analyses of pre-Q Comp data in order to inform the SRD substantive model specification and evaluate the merits of the SRD approach more broadly.
3. Estimate Q Comp's GLATEs on student achievement over time by subject.
4. For Q Comp participants, estimate the impact of individual districts and schools on student achievement (i.e., value added).

The first step will involve checking for misallocation and excluding distal observations from the comparison group. Did any schools located in Q Comp districts not participate, or did any schools in non-participating districts participate? Misallocation rates will be reported, but because the intent-to-treat effect is of interest, no remedy will be applied. The distance exclusion criterion will be set at the maximum distance that a Q Comp school is located from the nearest Q Comp border at any time period. Applying this criterion will include all Q Comp schools, but comparison schools will only be included if they were located less than or equal to the distance of furthest Q Comp school.

The second step will use pre-Q Comp data to guide specification of the substantive model and to evaluate the SRD approach. Findings from the second step will be used to choose a functional form for the relationship between student achievement and distance from the Q Comp border. Findings will also reveal the degree to which pre-existing border discontinuities might threaten the validity of conclusions about Q Comp

had data from pre-test years not been available to include in the substantive model as planned (i.e., the warrant for a longitudinal difference-in-differences approach).

The third step will estimate Q Comp's effects on student achievement. Using findings from the pre-Q Comp analyses, a longitudinal SRD model will be specified to estimate initial GLATEs and change over time by subject (math and reading) relative to nonparticipants. Multilevel/mixed effects models will be specified to account for group dependencies and to conservatively estimate the value added by Q Comp districts and schools. Even though the substantive model (third step) will be informed by the pre-test analyses (second step), it is important to use prior theory to specify and describe a plausible substantive model in advance to help ensure the pre-test analyses are consistent with and capable of guiding the subsequent analyses.

Substantive SRD model of Q Comp's effectiveness

I have specified a plausible substantive SRD model that applies many of the enhancements described in Table 3.1 in order to evaluate Q Comp's effectiveness. The model is the centerpiece of this thesis. The plausible specification may change depending on findings from analyses of pre-Q Comp data. The model features a basic SRD portion, a longitudinal portion, and random effects for the multilevel/longitudinal data structure.

The basic regression discontinuity portion of the model specifies that math and reading scores vary in a piecewise fashion according to distance from the Q Comp border interacted with Q Comp participation. Under regression discontinuity theory, distance is the continuous assignment variable that determines participation when "cut" by a border. I assume, based on exploratory plots and guidance by Shadish et al. (2002), that the relationship between distance and student achievement is cubic. I also presume that

control variables should be included to adjust for selection/sorting and other factors beyond the control of schools—factors not ruled out by the SRD natural experimental design involving pre-existing borders.

The longitudinal portion of the model specifies that student achievement *also* varies over time in a piecewise fashion (i.e., not just piecewise with respect to distance from the border). It features two time variables: one indexing scores by year for all schools over all years (2004-2013) and one for Q Comp schools in participating years (i.e., years of participation). This specification qualifies as a between-schools interrupted time series quasi-experiment. Q Comp levy share is included in order to estimate the degree to which additional, locally levied revenue moderates program effectiveness. The model specifies random effects to account for nesting of observations within schools and districts and over time.

The plausible model specification is:

$$\begin{aligned} Score_{ijk} = {} & \gamma_{000} + \gamma_{100} Year_i + \gamma_{200} QComp_{ijk} + \gamma_{300} YearsParticipation_{ijk} \\ & + \gamma_{400} QCompLevy_{ijk} + \gamma_{500} Distance_{ijk} + \gamma_{600} Distance_{ijk}^2 \\ & + \gamma_{700} Distance_{ijk}^3 + \gamma_{800} (QComp_{ijk} * Distance_{ijk}) \\ & + \gamma_{900} (QComp_{ijk} * Distance_{ijk}^2) + \gamma_{(10)00} (QComp_{ijk} * Distance_{ijk}^3) \\ & + \gamma_{(11)00} CensoredDistance_{ijk} + \gamma_{(12)00} Grade3_{ijk} + \gamma_{(13)00} Grade5_{ijk} \\ & + \gamma_{(14)00} GradeHS_{ijk} + \gamma_{(15)00} EnglishLearners_{ijk} \\ & + \gamma_{(16)00} Mobility_{ijk} + \gamma_{(17)00} Poverty_{ijk} + \gamma_{(18)00} Segregation_{ijk} \\ & + \gamma_{(19)00} SpecialEducation_{ijk} \\ & + r_{0jk} + r_{1jk} Year_i + u_{00k} + u_{10k} Year_i + e_{ijk} \end{aligned} \quad (3.1)$$

where:

- *Score* denotes school *j*'s (in district *k*) average of students' standardized math or reading test scores (i.e., mean z-score) in year *i*

- *Year* denotes the spring of calendar year *i* minus 2004, which is the first year of pre-Q Comp test scores to be included in the data set, so that *Year* = 0 for the 2003-2004 school year
- *QComp* indicates school *j*'s Q Comp participation in a given year (1 if yes; 0 otherwise), determined by district *k*'s participation unless misallocated
- *YearsParticipation* denotes the number of years that school *j* has participated in Q Comp, with the first year equal to 1
- *QCompLevy* denotes participating district *k*'s log odds of maximum allowable revenue levied on top of state Q Comp, centered on the grand weighted mean of schools in participating districts; set to zero if school *j* is misallocated
- *Distance* denotes school *j*'s distance to the Q Comp border in year *i*, centered on the border with positive values indicating location inside Q Comp region; set equal to -1 * the grand maximum distance of Q Comp schools if absolute distance is greater than the grand maximum
- *CensoredDistance* denotes a dummy variable equal to 1 in year *i* if a comparison group school located an absolute distance greater than the grand maximum distance of Q Comp schools; 0 otherwise
- *Grade3*, *Grade5*, and *GradeHS* denote grade level dummy variables equal to one for grade 3, 5, and 10 or 11, respectively; 0 otherwise and all 0 if grade 7.
- *EnglishLearners* denotes the log odds ratio of students who were English language learners, centered on the grand weighted mean

- *Mobility* denotes the log odds ratio of students who were not enrolled in the same school on October 1 of the school year, centered on the grand weighted mean
- *Poverty* denotes the log odds ratio of students eligible for free or reduced price lunch, centered on the grand weighted mean
- *Segregation* denotes the log odds ratio of students who identify as students of color and/or Hispanic or Latino ethnicity, centered on the grand weighted mean
- *SpecialEducation* denotes the log odds ratio of students receiving special education, centered on the grand weighted mean
- *r* and *u* denote random intercepts and slopes at the school and district levels, respectively
- *e* denotes random error.

Table 3.2 summarizes the key substantive evaluation questions and the corresponding fixed effects to be estimated.

Table 3.2. Summary of evaluation questions and corresponding fixed effects

Research question	**Parameter**
To what degree has Q Comp impacted student achievement in the first year of a schools participation (schools levying average of allowable share)?	γ_{200}
To what degree has Q Comp influenced student achievement over time?	γ_{300}
To what degree has Q Comp levy revenue moderated the program's impact on student achievement?	γ_{400}

Proportions are transformed to log odds for a couple reasons. Student characteristic proportions are first analyzed as outcomes in the first set of pre-test models, and one is later treated as a nonequivalent dependent variable (see Equation 3.4). They are tranformed to log odds ensure that predictions fall within the possible bounds (similar

to logistic regression). They are later placed on the right hand side as controls. Rather than convert them back to proportions, they remain as log odds on the right hand side 1) for consistency with the first set of pre-test models and 2) because exploratory plots suggest that log odds simplify further transformations (i.e., polynomials) for explaining student outcomes. That is, plotting student outcomes on proportions suggested slightly more complex methods (e.g., splines) might be needed to fit abrupt changes in form. Student log odds are also centered so that the estimated intercept pertains to a typical school (as opposed to schools with 50 percent of students each category) and to reduce collinearity between lower- and higher-order terms in a polynomial. Reducing collinearity will in turn deflate standard errors and reduce type II error.

As discussed by Imbens and Lemieux (2007), an instrumental variable approach could help alleviate bias introduced by misallocation, but estimating the intent-to-treat effect is preferable for program evaluation purposes (as long as misallocation rates are somewhat low) and because instrumental variable approaches entail additional costs. Missallocation can be thought of as non-compliance with treatment assigned by a regression discontinuity design that adds measurement error to and attenuates estimates of the true LATE. Similarly, unobserved iterations and levels of intensity of the professional development component of Q Comp can also be thought of as non-compliance (i.e., with competing Q Comp designers' ideal types, similar to the "manipulation" validity threat). Misallocation and other types of noncompliance can and should be expected in future iterations of Q Comp and similar programs. Therefore, the attenuation of Q Comp estimates actually serves a validity purpose: it tempers estimates and generalizability of the true treatment-on-treated effects of teacher improvement

systems, which are not possible to implement with full control and fidelity. If MDE and similar authorities remediate noncompliance, then estimates of Q Comp and similar programs' effectiveness would increase, all else being equal. Lastly, Wong, Steiner, and Cook (2010) found that applying an instrumental variable approach entailed costs without adding any benefits to their multivariate regression discontinuity study.

Value-added model

Which Q Comp districts and schools added significant value (i.e., in which Q Comp sites did students exhibit the largest achievement gains)? A feature of the Q Comp program is that it is not overly prescriptive. That is, MDE has permitted school districts to propose and implement a range of strategies within the parameters of the law and rules (Office of the Legislative Auditor, 2009; Schwartz, 2012). Additionally, variations in Q Comp implementation have arisen to some degree from a lack of clarity about program requirements. The various approaches have not been fully documented, making it difficult to identify which strategies are more effective than others. Identifying exemplary districts and schools that "added value" to student achievement would allow MDE or others to apply qualitative inquiry to identify strategies that should be emulated statewide to improve the Q Comp program overall.

The value-added model qualifies as an interrupted-time-series-as-outcomes model:

$$\begin{aligned} Score_{ijk} - \widehat{Score}_{ijk} = {} & \gamma_{000} + r_{0jk} + r_{1jk} QComp_{ijk} + r_{2jk} YearsParticipation_{ijk} \\ & + u_{00k} + r_{10k} QComp_{ijk} + r_{20k} YearsParticipation_{ijk} + e_{ijk} . \end{aligned} \quad (3.2)$$

Two features distinguish this model from the model of Q Comp's overall impact on student achievement (see Equation 3.1).

1. The dependent variable is residualized student achievement (i.e., observed values minus values fitted by the estimated fixed effects). Since the prior model specified fixed effects for Q Comp and years of participation, those parameters are assumed to be zero *between* schools and districts in the value-added model but not *within*.
2. The sample includes only schools that participated in Q Comp at any point in time (i.e., no comparison group schools, whether correctly allocated or misallocated within a Q Comp district).

The random Q Comp intercepts (i.e., r_{1jk} and u_{10k}) represent initial interruptions and the random Q Comp slopes (i.e., r_{2jk} and u_{20k}) represent slope interruptions. The random effects for Q Comp and years of participation are not included in the substantive model (i.e., the overall and value-added models were not combined) because the comparison group accounted for about half of the schools in the sample. That is, because the treatment and duration random effects are only applicable/defined for participating schools, including them would bias their variance component estimates downward.

The value-added model allows each school and district's random effects to be *predicted* (Snijders & Bosker, 1999). That is, their observed effects (and standard errors) are assumed to be measured with error and thus are proactively regressed toward the mean as a function of reliability. The resulting values qualify as "true/universe" scores (and standard errors of estimation) under classical/generalizability test theory. Exemplary schools are those that exhibited significant student achievement interruptions

of their own pre-Q Comp achievement levels. That is, they stand out because of one of the following:

1. student achievement exhibited an initial increase that subsequently remained steady or increased over time
2. student achievement did not exhibit an initial increase but subsequently increased over time.

Predictions are considered statistically significant if they are two or more standard errors from zero.

Pre-test SRD models

I propose analyzing student data from time periods prior to Q Comp (i.e., from 2004-2005) in order to guide substantive model specification and evaluate the SRD approach. I have specified two sets of pre-test models that resemble the substantive model and covariate analyses by Black (1999) and Bayer et al. (2007). The first set of models is largely exploratory and addresses the question, "To what degree were Q Comp schools similar to non-participating schools at the border in terms of student achievement and other characteristics?" In other words, does the SRD approach yield locally well-matched/balanced counterfactuals at the cutoff beforehand? Given that Q Comp qualifies as an SRD natural experiment due to pre-existing borders, it is quite possible that schools at the border are not ignorably different. The second set of models addresses the question, "Controlling for student characteristics, to what degree were Q Comp schools similar to non-participating schools at the border in terms of student achievement?" The second set of models will be used to refine the plausible specification of the substantive model.

For both sets of pre-test models, I will systematically vary distance from the Q Comp border to find an optimal combination of its operational definition and functional form. The operational definitions are Geographic Minimum distance, Geographic Centroid distance, Habitation Minimum distance, and Habitation centroid distance. The functional form conditions are linear through quartic. Crossing the operational definitions with the functional forms results in 16 total conditions per student dependent variable. An optimal combination is one that results in student characteristic balance at the Q Comp border and good model fit.

Operationalizing distance

The operational definitions can be thought of as vectors, each possessing a magnitude of a particular scale and a direction from a school's location to a Q Comp border (see Table 3.3). Computing vector termini (i.e., where lines meet polygons) was accomplished with the *spatstat* package (Baddeley & Turner, 2005) in *R* (R Core Team, 2014). The geographic scale in this study is kilometers (i.e., Universal Trans Mercator zone 15 meters divided by 1000). Geographic Minimum distance is probably the most familiar type (i.e., "as the crow flies"), with direction determined by the angle that minimizes the length of the vector from the school to nearest Q Comp border. Geographic Centroid distance is also measured in kilometers, but direction is not minimized. Rather, it is determined by the centroid of Q Comp district polygons. For a school in a participating district, the vector radiates away from the centroid, starting at the school, and measured to the border. For a school located outside of the district, the vector is measured from the school directly toward the centroid, stopping where it meets the border.

Table 3.3. Distance vector conditions

Determinant of direction	Magnitude scale	
	Geography	**Habitation**
Minimum distance	• Magnitude in kilometers • Direction determined by minimum distance	• Magnitude in standard deviations • Direction determined by minimum distance
Centroid	• Magnitude in kilometers • Direction determined by geographic centroid	• Magnitude in standard deviations • Direction determined by bivariate centroid

The scale for Habitation distances are not fixed statewide. Rather, the population density for each Q Comp district and it contiguous neighbors is used to construct a regional coordinate system with the goal of more closely approximating how humans perceive distance in social terms (e.g., distance traveled regularly for school or to purchase goods and services). For example, students and residents in the Twin Cities metropolitan area do not have to travel as many kilometers as those in rural Minnesota. Address clusters (i.e., midpoints of road line segments with house number ranges) from the 2010 Decennial Census (U.S. Census Bureau, 2012) are used to estimate the variance/covariance of residential locations in each Q Comp-plus-neighbors region. Next, eigenvectors are used to place school and border coordinates on standardized principal component scales (i.e., standard deviations from the two principal axes). Habitation Minimum distance is that which minimizes standard deviations from a school to the nearest Q Comp border. Habitation Centroid distance measures standard deviations either directly away from or directly toward the habitation centroid. Habitation centroid coordinates are zero on the habitation scale, which is the mean Easting and Northing of regional residences on the geographic scale. (Note that if one were interested in school Mahalanobis distances, then those values would be measured

from the habitation centroid all the way to sites on the habitation scale, not stopping at the Q Comp border.) Habitation Centroid distance may not be defined for a comparison group school if a region's habitation centroid lies outside of the Q Comp district polygon.

Figures 3.1 and 3.2 illustrate the distance typology. Schools are denoted *s* in the top-left panel. The geographic and habitation centroids are denoted *G* and *H*, respectively. The lighter, solid arrows in the right two panels represent minimum distance from each school to the nearest Q Comp border. The darker, dashed arrows represent centroid distance. Only Q Comp districts are labeled. Marshall and its neighbors offer a simple example (see Figure 3.1). It is a rural region with sparse population clusters. When Marshall participated in Q Comp in 2006, no nearby districts participated. Distance is defined for all schools under all four definitions because both the geographic and habitation centroid lie within Marshall's district border. Hopkins (Figure 3.2) is a more complex example. It is a first-ring suburban district with a greater density of residences and schools. When it participated in Q Comp in 2006, a contiguous district (Minneapolis) and some nearby districts also participated. The habitation centroid lies *outside* of Hopkins' border. Several schools in contiguous districts are either located in or closer to other Q Comp districts, which reflects another step in the final distance calculations. Namely, many distances may be computed for a school initially (e.g., if a district abuts more than one Q Comp district), but only the minimum distance is retained for later analysis.

Figure 3.1. Illustration of geographic and habitation scales: Rural district with no other Q Comp participants nearby

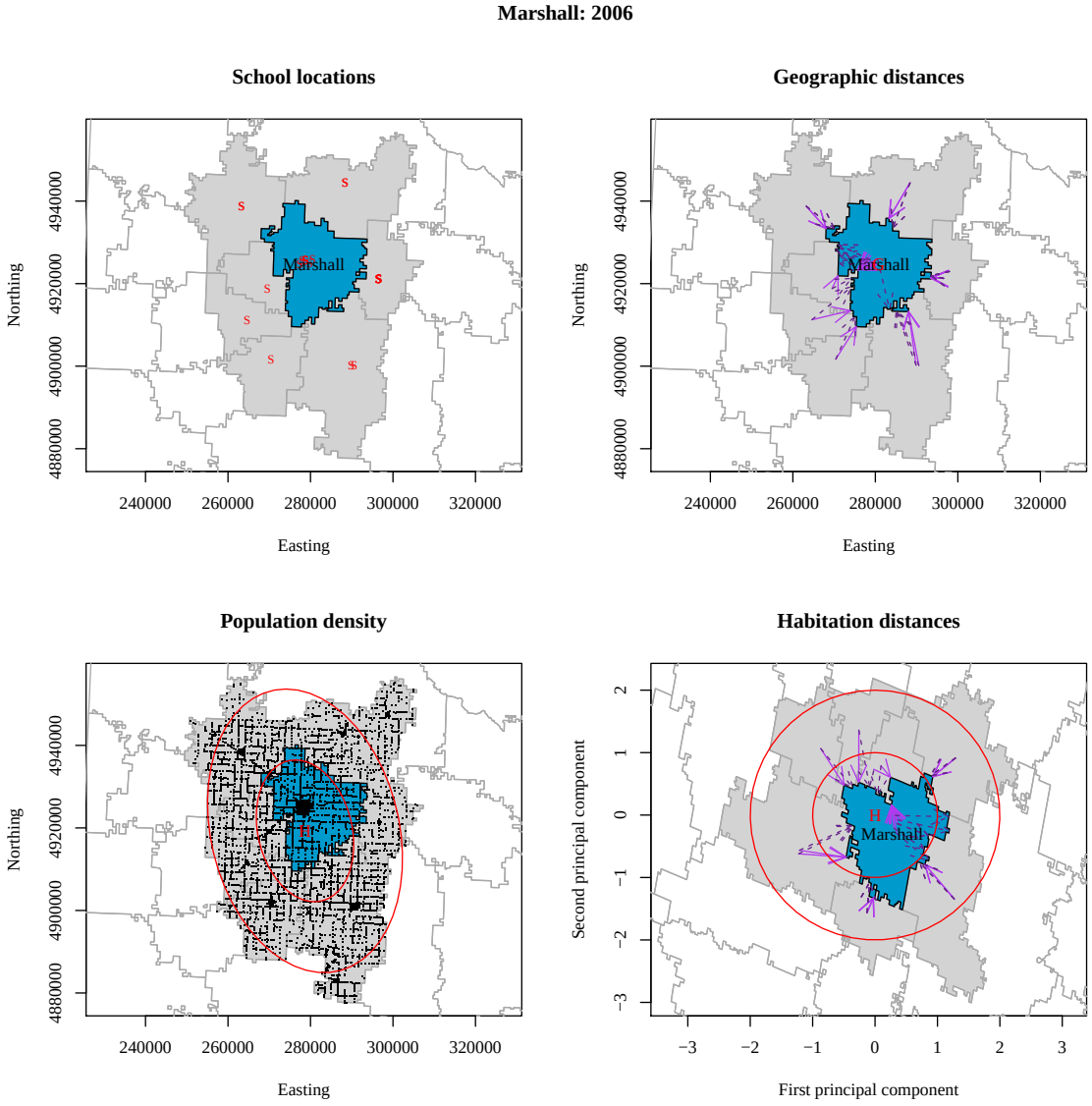
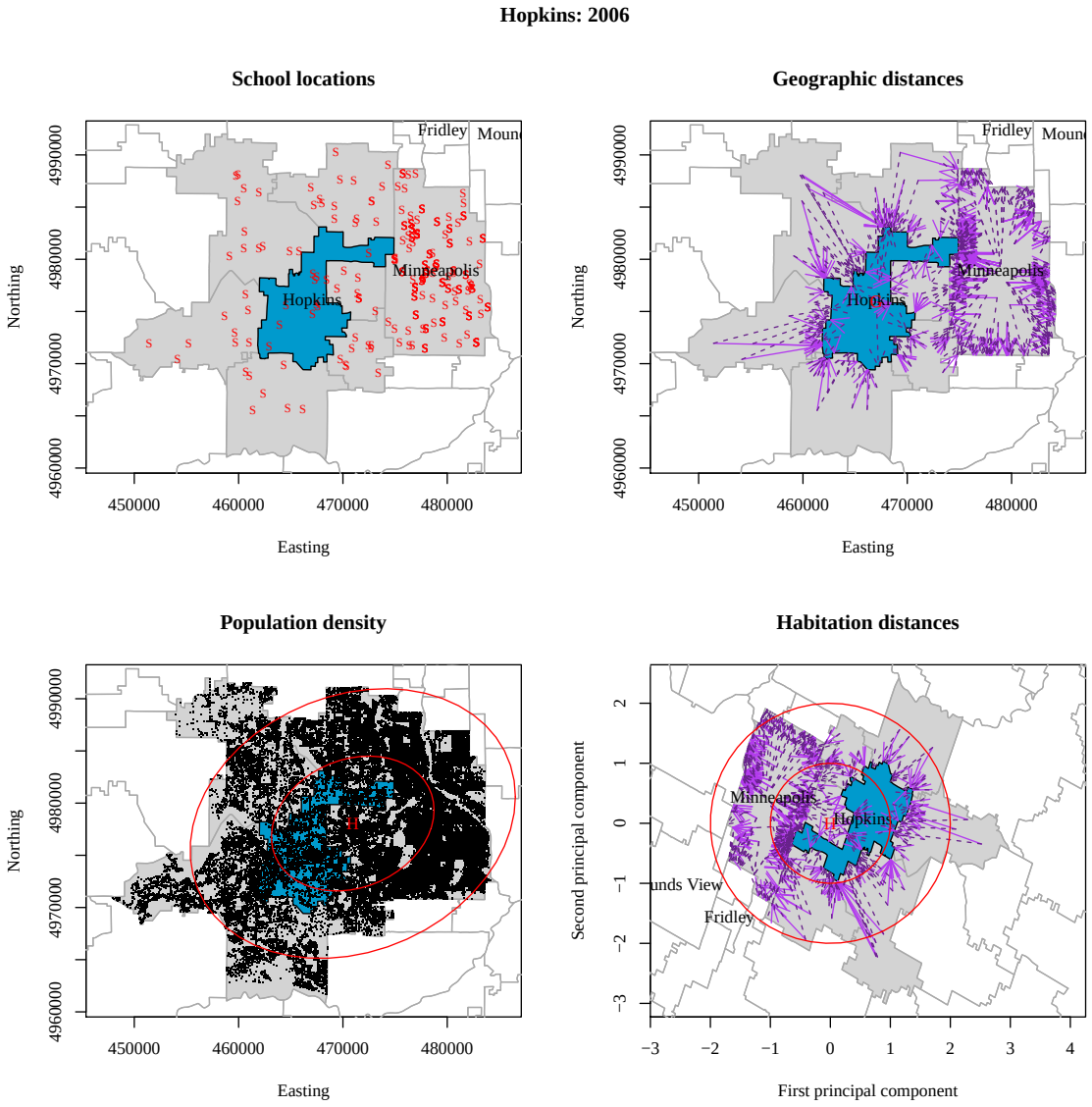


Figure 3.2. Illustration of geographic and habitation scales: Urban district with other Q Comp participants nearby



Pre-test SRD models of local student achievement and covariate balance

The first set of pre-test models involves regressing student achievement and characteristics prior to Q Comp on distance and participation after Q Comp began. The cubic specification is:

$$\begin{aligned}\overline{Y}_{jk} = {} & \gamma_{00} + \gamma_{10}\overline{QComp}_{jk} + \gamma_{20}\overline{Distance}_{jk} + \gamma_{30}\overline{Distance}^{2}_{jk} + \gamma_{40}\overline{Distance}^{3}_{jk} \\ & + \gamma_{50}(\overline{QComp}_{jk} * \overline{Distance}_{jk}) + \gamma_{60}(\overline{QComp}_{jk} * \overline{Distance}^{2}_{jk}) \\ & + \gamma_{70}(\overline{QComp}_{jk} * \overline{Distance}^{3}_{jk}) + \gamma_{80}\overline{DistanceCensored}_{jk} + u_{0k} + e_{jk} \quad ,\end{aligned} \tag{3.4}$$

where:

- $\overline{Y}$ denotes school *j*'s (in district *k*) two-year, pre-Q Comp average for a given dependent variable (students' standardized math or reading test scores and log odds of mobility, English learners, poverty, segregation, and special education)
- $\overline{QComp}$ denotes school *j*'s multi-year, post-Q Comp average participation
- $\overline{Distance}$ denotes school *j*'s multi-year, post-Q Comp distance to the Q Comp border
- *u* denotes random district intercepts
- *e* denotes random error.

The earliest, continuously participating Q Comp districts will have a $\overline{QComp}$ value of 1; late-adopters will have a $\overline{QComp}$ value between 0 and 1 (e.g., 0.14 for a school that started in the 2011-2012 school year); and $\overline{QComp}$ will equal 0 if a district has never participated. For schools in late-adopting districts, it is possible for $\overline{Distance}$ to have a negative value as a result of averaging and contrary to the deterministic assumption of the regression discontinuity design. As such, $\overline{QComp}$ values greater than 0 will be set to 0 if $\overline{Distance}$ is negative.

The coefficient $\hat{\gamma}_{10}$ is an estimate of pre-test balance at the Q Comp border for a given student characteristic (i.e., an estimate of the GLATE). Estimates of the other parameters will tell whether and to what degree a characteristic varies geographically with distance from the border. If there is enough evidence to conclude that student scores

and characteristics vary with distance from the border and if balance improves with proximity to the border, then that would lend support for the SRD approach.

Pre-test SRD models of local student achievement balance

The second set of pre-test models involves regressing student achievement prior to Q Comp on distance, participation after Q Comp began, and student covariates. Note that non-achievement student characteristics are now on the right-hand side of the model, acting as control variables. The cubic specification of the second set of pre-test models is:

$$\begin{aligned}\overline{Score}_{jk} = {} & \gamma_{00} + \gamma_{10}\overline{QComp}_{jk} + \gamma_{20}\overline{Distance}_{jk} + \gamma_{30}\overline{Distance}^{2}_{jk} + \gamma_{40}\overline{Distance}^{3}_{jk} \\ & + \gamma_{50}\left(\overline{QComp}_{jk} * \overline{Distance}_{jk}\right) + \gamma_{60}\left(\overline{QComp}_{jk} * \overline{Distance}^{2}_{jk}\right) \\ & + \gamma_{70}\left(\overline{QComp}_{jk} * \overline{Distance}^{3}_{jk}\right) + \gamma_{80}\overline{DistanceCensored}_{jk} \\ & + \gamma_{90}\overline{EnglishLearners}_{j} + \gamma_{(10)0}\overline{Mobility}_{j} + \gamma_{(11)0}\overline{Poverty}_{j} \\ & + \gamma_{(12)0}\overline{Segregation}_{j} + \gamma_{(13)0}\overline{SpecialEducation}_{j} \\ & + u_{0k} + e_{jk} \end{aligned} \quad (3.5)$$

.

The coefficient $\hat{\gamma}_{10}$ is an estimate of pre-test student achievement balance at the Q Comp border (i.e., an estimate of the GLATE). Estimates of the other parameters will tell whether and how student achievement varies geographically and with other student characteristics. With control variables accompanying the assignment and treatment variables (i.e., distance and Q Comp), it is possible that schools will exhibit comparable student achievement at the Q Comp border, even if that was not the case without controls (i.e., in the first set of pre-test models). If local student achievement balance is not achieved even after controlling for student characteristics, then that would represent nonignorable evidence of selection, although the substantive model specification preemptively addresses that possibility by including pre-Q Comp years.

Estimation

Maximum likelihood estimates of the parameters will be obtained using the *lme4* package (Bates et al., 2014) in *R* (R Core Team, 2014). Larger schools will be given more weight during estimation to help ensure that Q Comp inferences generalize to students more broadly than to schools. Schools will be weighted by the square root of the number of students tested and standardized to sum to 1 when employed in model estimation. Intuitively, this will avoid giving disproportionately large influence to small schools (if unweighted) and avoid giving extreme weight to extremely large schools (i.e., taking the square root effectively pulls in the positively skewed distribution of raw counts of students tested). Perhaps most importantly, the proposed weighting scheme approximates inverse-variance weighting in which schools with more precisely measured mean z-scores are given optimal weight during estimation.

Model selection and inference strategy

A model selection strategy will be applied to choose among the many pre-test models. I have proposed estimating balance and fit for two student achievement outcomes in two ways (with and without controls: pre-test model sets 1 and 2) as well as five other student characteristics (pre-test model set 1) for a total of nine base models. Crossing those models with the four distance definitions and four functional forms yields a total 144 pre-test exploratory models. An optimal model is one that results in balance at the Q Comp border and good model fit. GLATE estimates of pre-Q Comp student characteristics represent balance. Fit will be quantified as the proportion of variation explained by the fixed effects, a pseudo-R^2 measure appropriate for multilevel models (Long, 2012). The formula, adapted for the student-weighting strategy in this study, is:

$$R^2_{(y\,\hat{y})} = \left(\frac{\dfrac{\sum w_{ijk}\left(y_{ijk} - \dfrac{\sum w_{ijk}\,y_{ijk}}{\sum w_{ijk}}\right)\left(\hat{y}_{ijk} - \dfrac{\sum w_{ijk}\,\hat{y}_{ijk}}{\sum w_{ijk}}\right)}{\sum w_{ijk}}}{\sqrt{\dfrac{\sum w_{ijk}\left(y_{ijk} - \dfrac{\sum w_{ijk}\,y_{ijk}}{\sum w_{ijk}}\right)^2}{\sum w_{ijk}}}\sqrt{\dfrac{\sum w_{ijk}\left(\hat{y}_{ijk} - \dfrac{\sum w_{ijk}\,\hat{y}_{ijk}}{\sum w_{ijk}}\right)^2}{\sum w_{ijk}}}} \right)^2, \quad (3.6)$$

where y is a student dependent variable, $\hat{y}$ is the value predicted by the fixed effects, and w denotes the square root of the number of students at time *i* in school *j* in district *k*. Observations are not time specific in the pre-test analysis, but they are indexed by time *i* in the final models of Q Comp's influence on student achievement. Averaging over time reduces variability in student characteristics and thus may inflate R^2 values in the pre-test analysis relative to the final models.

The model selection strategy will whittle down the 144 pre-test specifications to two final fixed effects specifications for estimating Q Comp's influence on student achievement. The strategy will involve plotting balance against model fit. The optimal model(s) will appear closest to zero on the y-axis and closest to one on the x-axis (i.e., smallest absolute GLATE and largest R^2). The optimal model(s) will guide specification of the substantive model, resulting in a final model that accounts for validity threats and is more parsimonious than what could be achieved without pre-test analyses. Additionally, the plot of balance against fit will reveal the degree to which SRD natural experiments are warranted (i.e., the degree to which balance is both local and sufficient).

Estimates of the parameters listed in Table 3.2 will be interpreted in terms of their significance and standardized effect size. The identities of schools and districts that

added significant value will be reported and discussed to help inform potential improvements to the Q Comp program. The Minnesota Department of Education (MDE) could use those findings to better identify program characteristics that are effective and scale them up to improve Q Comp's overall effectiveness.

Checking assumptions

The validity of statistical conclusions from multilevel models depends on meeting key assumptions that will be checked during the analysis phase of this study (Snijders & Bosker, 1999). Ordinary least squares (OLS) regression assumes that the outcome is linearly dependent on the explanatory variable(s) and that important variables are not missing from the right-hand side of the equation. This assumption is extended to both fixed and random effects at every level of the multilevel model. With regard to residuals, OLS assumes that they are normally distributed with constant variance. This assumption still applies at level one of the multilevel model and is extended to the random coefficient covariance matrix. As discussed throughout this thesis up to this point, great care has been taken to specify models that, based on prior theory, available variables, and pre-test analyses, will lead to valid conclusions about Q Comp. Bivariate relationships will be explored in greater detail prior to fitting each set of models to help ensure that important variables and transformations have not been overlooked. Normality and homoscedasticity of residuals and random effects will be checked as models are fit.

Districts and schools: Fixed or random?

This study treats districts and schools as random effects in a mixed effects model for the following reasons: 1) to avoid ecological fallacies, 2) to obtain design-based standard errors, 3) to conservatively identify exemplary districts and schools, and 4) for

consistency with how Q Comp was designed, implemented, and documented. Mixed effects modeling shares many similarities with classical psychological test theory and generalizability theory (Snijders & Bosker, 2012; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001). That is, basic models treat level-one observations as representative indicators of an indirectly observable latent factor (i.e., measured with error). Mixed effects modeling helps ensure valid inferences by regressing a dependent variable on 1) level-two (or higher) latent factors, 2) observed/manifest fixed effects and 3) random effects (i.e., grouping variables). That is, mixed effects modeling helps avoid ecological fallacies (e.g., Simpson's paradox) and inflation of Type I error that can occur naively, when within-group dependencies arising from multistage sampling are ignored. It avoids the former by essentially performing regressions at each level/stage of the sample (i.e., within and between groups). It avoids the latter in two ways: 1) by adjusting the effective sample size for multistage sampling (i.e., downward from the total number of level-one observations) and 2) by conservatively shrinking group effects toward the population mean to the degree that a group size is small and the group effect not reliable (i.e., by "estimating" true/universe scores instead of observed scores). Gelman, Hill, and Yajima (2012) contend that shrinking is conservative and appropriate because it qualifies as a Bayesian alternative to Bonferroni and other multiple comparison corrections.

Mixed effects modeling is not necessary to avoid problems from naively modeling clustered data. Indeed, Q Comp studies by Schwartz (2012) and Sojourner et al. (in press) treat level-two observations (schools in the former study and students in the latter) as fixed effects (i.e., dummy variables) and then calculate standard errors that are robust to nesting of observations within districts (i.e., without including district dummy

variables in the model). Additionally, true/universe scores can be obtained by shrinking fixed effects that are not 100 percent reliable (Herrmann, Walsh, Isenberg, & Resch, 2013; Value-Added Research Center, 2013). Snijders and Bosker (2012) note that there is no single, clear-cut reason for treating grouping variables as random effects instead of fixed effects, and Gelman (2005) notes that "fixed effects" and "random effects" carry several different and conflicting meanings. They suggest simply treating grouping variables as fixed effects if they represent theoretically distinct and/or manipulated levels for inference (e.g., treatment and control) and random effects if they qualify as clusters sampled from a larger population that help explain additional variation left over by fixed effects (i.e., by allowing effects to vary from one group to another).

I contend that districts and schools should be treated as random effects because their influence on student achievement represents the indirectly observed instructional improvement component of Q Comp. As discussed by Schwartz (2012) and Sojourner et al. (in press), Q Comp qualifies as a voluntary grantor-grantee teacher improvement program with an incentive component enforced by MDE and an instructional improvement component implemented by districts and schools. Combining evaluation and measurement perspectives (as mixed effects modeling does), student achievement is the latent outcome of interest, Q Comp qualifies as the treatment, and selection allows construct irrelevant variance (i.e., that cannot be balanced experimentally). The incentive component of Q Comp is directly observable because MDE took on that role and documented compliance. However, the instructional improvement component is not directly observable. Its implementation was diffuse, under the purview of districts and schools. This was done partly by design (i.e., to encourage participation and innovation)

and partly due to MDE's limited capacity both in terms of expertise and budget. In other words, the instructional improvement component qualifies as a latent factor measured with error and that is only indirectly/reflectively observable by manifest district and school effects. District and school effects may also capture effective applications of the incentive component at the local level. As discussed above, Schwartz (2012) and Sojourner et al. (in press) made use of grouping variables to help ensure valid inferences, but they did not attempt to quantitatively identify exemplary Q Comp districts and schools that added value to student learning. Figure 3.3 shows which combinations of variables correspond to which sources of variation, whether variation can be observed directly or indirectly, and estimation decisions stemming from the measurement scenario.

Figure 3.3. Conceptual variance decomposition, given Q Comp's design/implementation and observables

Theory	Construct	Variable	Estimation
Student academic achievement =	Outcome	Latent	Weighted
Student characteristics + *Context effects* +	Irrelevant: Imbalance	Observed	Fixed
Law + *Rule* +	Q Comp: Incentives and effort	Observed	Fixed
District + *School*	Q Comp: Instructional supports and ability	Latent	Random

Summary of methods

In summary, I propose fitting four sets of multilevel/longitudinal SRD models: two sets of pre-Q Comp models of student achievement and characteristics before the program began (i.e., from 2004-2005); a substantive set of models of student achievement from before and after Q Comp began in 2006 (i.e., from 2004-2013); and a

set of value-added models. The primary purpose of the pre-Q Comp models is to guide the substantive models in terms of 1) how distance should be operationalized and 2) the functional form of the relationship between distance and achievement. The pre-Q Comp models will also provide insights about the degree to which the SRD approach yields well-matched/ignorably different schools at district borders in terms of student achievement and covariates. The substantive models will provide estimates of Q Comp's impact on student achievement, as well as the degree to which the program's additionally levied revenue has moderated its effectiveness. Additionally, the value added by each Q Comp school and district will be predicted so that those with exemplary outcomes can be identified and emulated. The value-added step is important because Q Comp is not prescriptive, especially with regard to the instructional support component, and some schools and districts may have implemented Q Comp in ways that proved more effective than others. Taken together, the proposed methods demonstrate how SRD and the best practices identified in Chapter 2 can be applied to a program that has the potential to improve teaching and student achievement but that others have found difficult to evaluate in the past. Table 3.4 summarizes the proposed methods.

Table 3.4. Summary of methods

Model	Purpose	Specifications/conditions	Model selection and inference criteria
Pre-test	• Guide substantive model specification • Evaluate the SRD approach	• Specifications ◦ Set 1: Scores and covariates regressed on treatment and assignment ◦ Set 2: Scores regressed on treatment, assignment and covariates • Conditions ◦ Geographic Minimum, Geographic Centroid, Habitation Minimum, and Habitation Centroid distance ◦ Linear, quadratic, cubic, and quartic functional forms for distance	• Balance: GLATE estimate • Fit: Weighted pseudo-R^2
Substantive	• Estimate Q Comp's impact on student achievement • Estimate the moderating influence of levied revenue	• Plausible specification ◦ Scores regressed on treatment, distance, and covariates	• Statistical significance • Standardized effect sizes
Value added	• Identify exemplary Q Comp schools and districts	• Residualized achievement regressed on intercept and slope interruptions	• True scores and confidence intervals

Chapter 4: Results

Introduction

This chapter presents results from the spatial regression discontinuity (SRD) models. Chapter 2 integrated literature pertaining to quasi-experimental methods and the validity of geographically local average treatment effect (GLATE) estimates from SRD designs. Chapter 3 proposed applying best practices to evaluate Minnesota's Quality Compensation for Teachers (Q Comp) program. The pre-test models explore balance at the Q Comp border and how distance is related to student characteristics. The substantive and value-added models incorporate findings from the pre-test models in order to estimate Q Comp's impact on student achievement and identify exemplars.

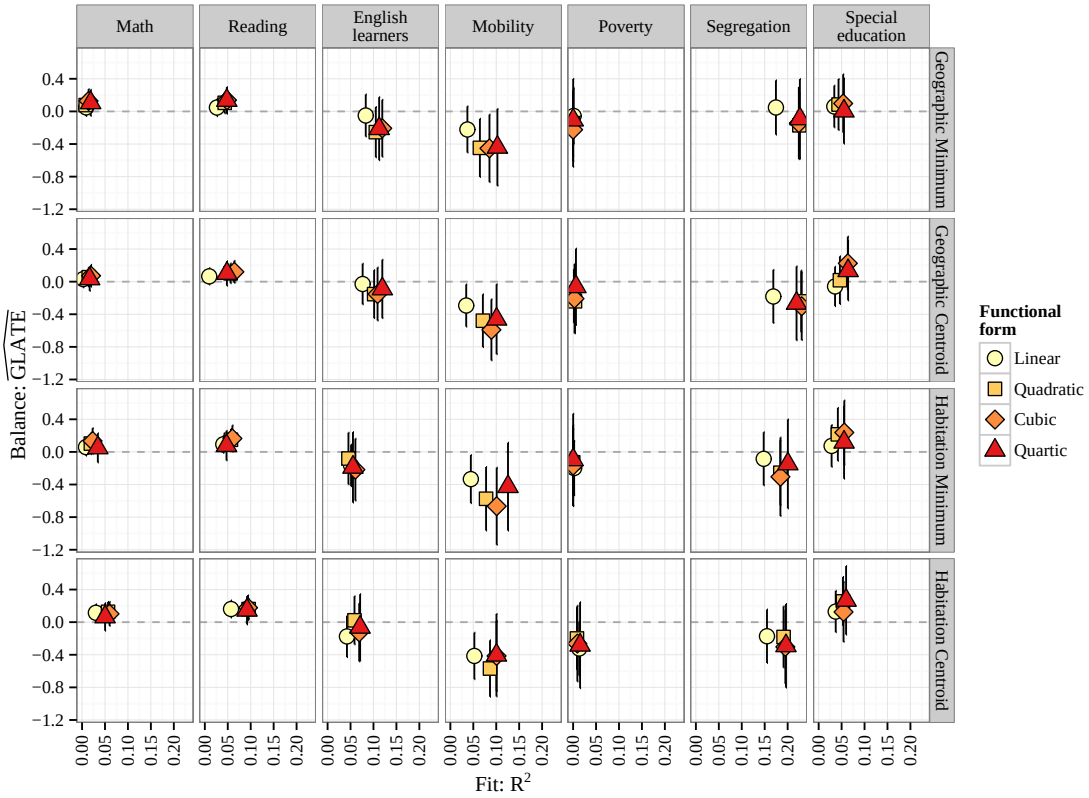
Exploratory analysis of fit and functional form

Pre-test model set 1

In the first set of exploratory models, piecewise distance explains a small amount of variation in math and reading achievement (see Figure 4.1). For math, distance explains from less than 1% to about 6% depending on how distance is operationalized and the functional form. For reading, distance explains about 1% to 10%. In most instances, the models do not provide enough evidence to reject the null hypothesis that math and reading achievement were equal at the Q Comp border before the program began, as evidenced by the confidence intervals encompassing zero in the plot. However, local math imbalance can be inferred in two instances (12.5%): when the Habitation Centroid definition is used in combination with either a linear or quadratic functional form. For reading, local imbalance can be inferred in five instances (about 31%): the

quadratic and cubic specifications of Habitation Minimum distance and the linear, quadratic and cubic specifications of Habitation Centroid distance.

Figure 4.1. Estimates of balance and fit: Pre-test model set 1



Piecewise distance tends to explain more variation in the other, non-achievement student characteristics, with the exception of poverty. The models of segregation explain about 15%-23% of variation, depending on how distance is operationalized and specified. The models of English learners, mobility, special education, and poverty explain about 4%-12%, 3%-13%, 3%-6%, and <1%-2%, respectively. Differences in mobility at the Q Comp border can be inferred in 75% of instances (12/16). Local balance was much more evident in the models of English learners, poverty, segregation, and special education, with the only exception being poverty modeled by linear Habitation Centroid distance.

Model set 1 is highly exploratory, but some trends emerge. Student achievement and other characteristics tend to be similar at the Q Comp border, although inferences about balance are somewhat sensitive to the operational definition and functional form for distance. Adding higher-order terms for distance does not tend to substantially improve fit (i.e., the more parsimonious models suffice). Student mobility is an exception in two regards: 1) non-participating schools near the Q Comp border tended to have a higher share of students who arrived after October 1 than participating schools and 2) fit improved slightly with the addition of higher-order terms. This suggests that inheriting students from other schools may have discouraged selection into Q Comp. Like mobility, more variation in segregation can be explained with higher-order terms, but segregation is robustly similar/balanced at the Q Comp border.

In terms of an operational definition of distance, a clear winner does not emerge. That is, model fit does not vary greatly from one definition to another, but Habitation Minimum distance exhibits some advantages. Habitation Minimum distance improves homoskedasticity with respect to distance by taking district geographic area and population density into account when rescaling geographic coordinates on a region-wise basis (i.e., Q Comp districts and contiguous neighbors). That is, student achievement and other characteristics exhibit greater variability near the Q Comp border when distance is measured in kilometers, perhaps due to greater variability in urban areas. Variability appears more uniform when distance is in standard deviation units (after rescaling).

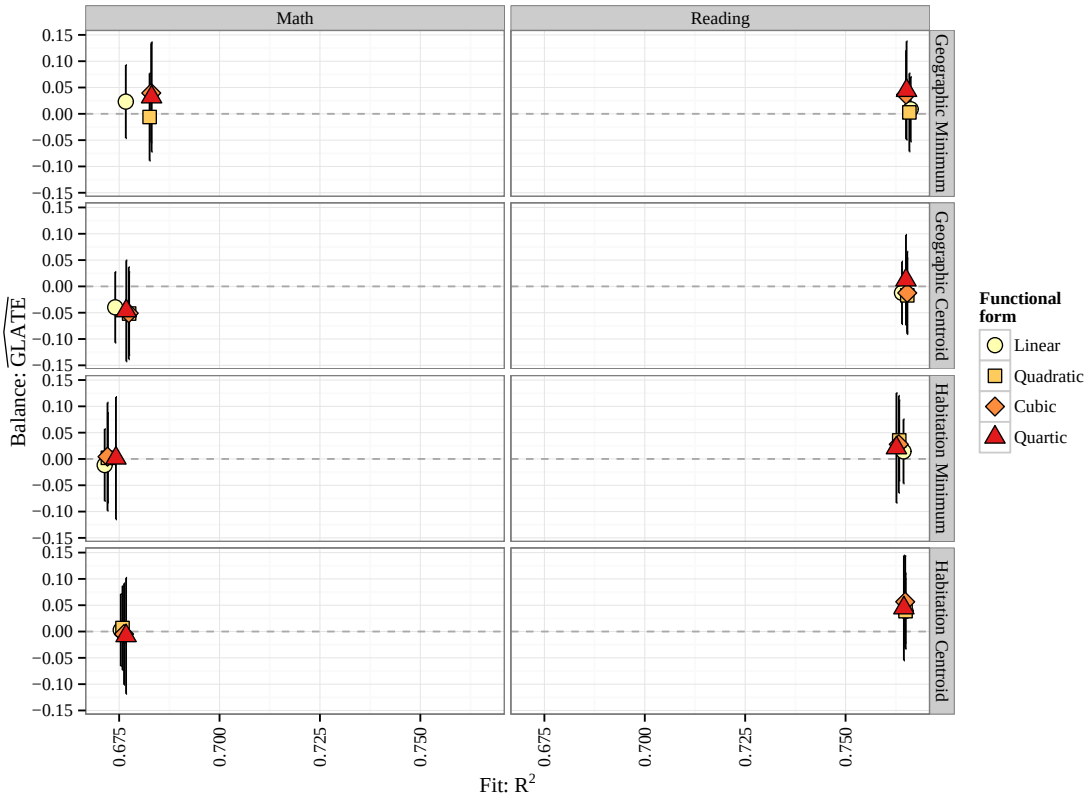
Pre-test model set 2

The second set of pre-test exploratory models regress pre-Q Comp math and student achievement on post-Q Comp piecewise distance while controlling for the non-

achievement student characteristics. These models more closely resemble the plausible model for evaluating Q Comp's impact on student achievement (see equation 3.1). Higher-order terms for the student control variables were added to approximate the functional form suggested by bivariate exploratory plots encompassing all time periods.

Results reveal that math and reading achievement were comparable at the Q Comp border before Q Comp was enacted (see Figure 4.2). Neither balance nor fit are sensitive to how distance is operationalized and specified. The models explain between 67%-68% of variation in math achievement and 76%-77% in reading, which represent large increases over the first set of pre-test models without controls. Since adding higher-order terms for distance does not substantially improve model fit, a parsimonious, linear model will suffice for estimating Q Comp's influence on student achievement. These findings of robust achievement balance and explanatory power at pre-test are highly desirable because they help rule out the fuzzy RD validity threat when post-test student achievement is used to evaluate Q Comp. That is, any changes in student achievement after participation are unlikely due to preexisting differences at the border.

Figure 4.2. Estimates of balance and fit: Pre-test model set 2



The balance and fit criteria do not reveal a clear winner for operationalizing distance, but Habitation Minimum has an important advantage. Geographic Minimum distance has the benefit of being the most familiar type (i.e., "as the crow flies"). The other operational definitions, from more familiar/simple/lenient to unfamiliar/complex/stringent are Geographic Centroid, Habitation Minimum and Habitation Centroid. Habitation distances carry a cost of being less understandable to general audiences because unlike geographic distance the scale depends on the population density of the region and not coordinates on the surface of the earth. However, rescaling geographic distance recognizes that student achievement is more variable in urban areas where the geographic area of the school districts is generally

smaller and the population density is greater. After rescaling, student achievement is more homoskedastic with respect to distance. Habitation Minimum distance has an advantage over Habitation Centroid distance in that the former does not require comparison group schools' distance vectors to pass through both the habitation centroid and the Q Comp border. That is, Habitation Centroid distance aggressively excludes comparison group schools if the habitation centroid does not lie within the Q Comp district, whereas minimum distance relaxes that requirement. Therefore, Habitation Minimum distance will be used going forward because it strikes a balance between applying a justifiable transformation and maintaining fidelity to the most familiar type of distance (Geographic Minimum). Taken together, results from the exploratory analysis of balance and fit help rule out the fuzzy RD validity threat and threats posed by inappropriately specifying an operational definition and functional form for the assignment variable.

Analysis of Q Comp's impact

Descriptive statistics

The Habitation Minimum models of Q Comp's impact on student achievement includes 21,035 cohorts overall (see Table 4.1). Cohorts are the units of analysis. A cohort is a group of students who took a Minnesota Comprehensive Assessment (MCA) or Mathematics Test for English Language Learners (MTELL) at the same grade level and in the same school in a given year. A cohort's mean of student z-scores is the outcome. In elementary through middle school, each cohort takes the math and reading MCA. In high school, a grade 10 cohort takes only the reading MCA, and a grade 11 cohort takes only the math MCA. As a result 17,975 cohorts have math scores and

17,868 have reading scores. Cohorts with less than 10 students are not included in the analysis because the data are not released by Minnesota Department of Education in order to protect student privacy. Table 4.1 shows descriptive statistics weighted by the square root of the number of students in each cohort. Note that values of 0 and 1 were treated as 0.025 and 0.975, respectively, when converting proportions to log odds because their values would be infinite otherwise.

Table 4.1. Descriptive statistics

Variable	N	Mean	SD	Minimum	Median	Maximum
Math	17975	-0.02	0.43	-2.48	0.04	1.66
Reading	17868	-0.02	0.40	-2.74	0.03	1.54
Year	21035	2008.47	2.87	2004.00	2008.00	2013.00
Q Comp	21035	0.23	0.42	0.00	0.00	1.00
Years of Q Comp participation	21035	0.81	1.78	0.00	0.00	8.00
Q Comp Levy	21035	-2.30	2.67	-3.66	-3.66	3.66
Habitation Minimum distance	21035	-0.31	0.54	-1.39	-0.12	1.39
Distance censored	21035	0.12	0.32	0.00	0.00	1.00
Grade	21035	6.65	3.01	3.00	7.00	11.00
English learners	21035	-3.01	1.09	-3.66	-3.66	3.66
Mobility	21035	-2.75	1.02	-3.66	-2.94	3.66
Poverty	21035	-0.93	1.41	-3.66	-0.90	3.66
Segregation	21035	-1.33	1.70	-3.66	-1.70	3.66
Special education	21035	-2.51	0.97	-3.66	-2.17	3.66
Students	21035	168.91	163.50	10.00	99.00	839.00

As discussed in Chapter 2, treatment misallocation can threaten the validity of results from a regression discontinuity study, and it is present in the current study. In a large majority of school districts that chose to participate in Q Comp, all schools participated. Three districts were exceptions: Minneapolis Public Schools, Roseville Public School District, and Alden-Conger Public School District. Contrary to assumption of deterministic treatment assignment, some schools located in those districts did not participate in Q Comp (see Table 4.2). About three percent of schools were misallocated

overall, with between four and six percent misallocated during the years that Q Comp was implemented. Given the low rates of misallocation and that it was permitted by both MDE and those districts, the misallocated schools remain in the analysis as comparison group schools but with a positive value for distance. All other schools have negative distance values if in the comparison group and positive distance values if in the Q Comp treatment group.

Table 4.2. School misallocation rates: Habitation Minimum distance

Year	Grade					
	3	**5**	**7**	**10**	**11**	**All**
2004	0.000	0.000	0.000	0.000	0.000	0.000
2005	0.000	0.000	0.000	0.000	0.000	0.000
2006	0.059	0.062	0.062	0.036	0.049	0.055
2007	0.060	0.061	0.062	0.053	0.044	0.057
2008	0.045	0.047	0.050	0.051	0.046	0.047
2009	0.041	0.045	0.048	0.034	0.048	0.043
2010	0.042	0.046	0.049	0.034	0.048	0.044
2011	0.038	0.040	0.042	0.042	0.044	0.041
2012	0.001	0.001	0.002	0.005	0.005	0.003
2013	0.001	0.001	0.002	0.005	0.004	0.003
All	0.029	0.030	0.032	0.026	0.029	0.029

Results

The final model specification is:

$$\begin{aligned} Score_{ijk} = {} & \gamma_{000} + \gamma_{100} Year_i + \gamma_{200} QComp_{ijk} + \gamma_{300} YearsParticipation_{ijk} \\ & + \gamma_{400} QCompLevy_{ijk} + \gamma_{500} Distance_{ijk} \\ & + \gamma_{600} (QComp_{ijk} * Distance_{ijk}) + \gamma_{700} CensoredDistance_{ijk} \\ & + \gamma_{800} Grade3_{ijk} + \gamma_{900} Grade5_{ijk} + \gamma_{(10)00} GradeHS_{ijk} \\ & + \gamma_{(11)00} EnglishLearners_{ijk} + \gamma_{(12)00} EnglishLearners^2_{ijk} \\ & + \gamma_{(13)00} Mobility_{ijk} + \gamma_{(14)00} Mobility^2_{ijk} + \gamma_{(15)00} Mobility^3_{ijk} \\ & + \gamma_{(16)00} Poverty_{ijk} + \gamma_{(17)00} Poverty^2_{ijk} + \gamma_{(18)00} Segregation_{ijk} \\ & + \gamma_{(19)000} Segregation^2_{ijk} + \gamma_{(20)00} SpecialEducation_{ijk} \\ & + \gamma_{(21)000} SpecialEducation^2_{ijk} + \gamma_{(22)000} SpecialEducation^3_{ijk} \\ & + r_{0jk} + u_{00k} + e_{ijk} \end{aligned} \quad (4.1)$$

Note that two changes were made to the initial model specification (Equation 3.1). Higher-order terms for student characteristics were added to account for curvilinear relationships observed in exploratory plots. Secondly, random slopes for years were backward eliminated due to variance components near zero and very high correlations with random intercepts returned by models that falsely converged. The lack of variation and high covariation can be attributed in part to the standardizing of the student achievement outcome within each year and to the saturation of the two observations in pre-test years for early adopters of Q Comp.

Tables 4.3.1 and 4.3.2 show results from the model of math achievement. The fixed effects explain about 56% of variation in math achievement. For a given observation, about 21% of unexplained variation is attributable to the district and about 40% to the school. Distance is not significantly different from zero on either side of the cutoff, and the distance fixed effects (i.e., piecewise distance and the censored distance indicator) explain very little additional variation (about 0.01%) in math achievement. This means that inferences about Q Comp's influence on math achievement are not geographically local. That is, they generalize to schools located in Q Comp and neighboring districts further from the border (but not necessarily to non-neighboring districts or highly distal schools, which were excluded from the sample). Additionally, the SRD method does not represent an improvement over the other quasi-experimental components employed in the model (i.e., distance could be backward eliminated for parsimony).

Table 4.3.1. Math results: Fixed effects

Fixed effect	Estimate	Standard error	t	p
Intercept	-0.0262	0.0159	-1.6469	0.0996
Year (centered on 2004)	0.0026	0.0007	3.4864	0.0005
Q Comp	0.0541	0.0120	4.5130	0.0000
Years of Q Comp participation	-0.0017	0.0017	-0.9580	0.3381
Q Comp Levy	-0.0072	0.0015	-4.9261	0.0000
Habitation Minimum distance	-0.0046	0.0075	-0.6212	0.5345
Q Comp * Habitation Minimum distance	-0.0527	0.0380	-1.3872	0.1654
Distance censored	-0.0050	0.0099	-0.5040	0.6142
Grade 3	0.0171	0.0088	1.9342	0.0531
Grade 5	0.0141	0.0085	1.6536	0.0982
Grade 11	-0.0256	0.0083	-3.0744	0.0021
English learners	-0.0008	0.0041	-0.1917	0.8480
English learners ^ 2	-0.0148	0.0019	-7.6565	0.0000
Mobility	-0.0747	0.0037	-20.0873	0.0000
Mobility ^ 2	-0.0582	0.0038	-15.2222	0.0000
Mobility ^ 3	0.0092	0.0006	15.7372	0.0000
Poverty	-0.0553	0.0025	-22.2489	0.0000
Poverty ^ 2	-0.0151	0.0009	-16.6940	0.0000
Segregation	-0.0513	0.0035	-14.7031	0.0000
Segregation ^ 2	-0.0019	0.0007	-2.6420	0.0082
Special education	-0.0356	0.0025	-14.4985	0.0000
Special education ^ 2	-0.0445	0.0040	-11.1813	0.0000
Special education ^ 3	0.0041	0.0007	5.5639	0.0000

Table 4.3.2. Math results: Random effects

Level	Random effect	Variance	Proportion
District	Intercept	0.0202	0.2126
School	Intercept	0.0377	0.3971
Time	Residual	0.0371	0.3903

Q Comp has positively and significantly impacted math achievement. Schools in Q Comp districts that levied the average amount exhibited 0.0541 standard deviation greater math achievement on average. There is not enough evidence to conclude that the effect changed over time. Q Comp levy share is significantly negatively associated with

math achievement, which decreases by -0.0072 standard deviation per allowable levy share (log odds) on average all else being equal.

Tables 4.4.1 and 4.4.2 show results from the model of reading achievement. The fixed effects explain about 61% of variation in reading achievement. For a given observation, about 22% of unexplained variation is attributable to the district and about 42% to the school. Distance is not significantly different from zero on either side of the cutoff, and the distance fixed effects (i.e., piecewise distance and the censored distance indicator) explain very little additional variation (about 0.12%) in reading achievement. This means that inferences about Q Comp's influence on reading achievement are not geographically local. That is, they generalize to schools located in Q Comp and neighboring districts further from the border (but not necessarily to non-neighboring districts or highly distal schools, which were excluded from the sample). Additionally, the SRD method does not represent an improvement over the other quasi-experimental components employed in the model (i.e., distance could be backward eliminated for parsimony).

Table 4.4.1. Reading results: Fixed effects

Fixed effect	Estimate	Standard error	t	p
Intercept	-0.0130	0.0139	-0.9384	0.3481
Year (centered on 2004)	0.0033	0.0006	5.2977	0.0000
Q Comp	0.0247	0.0101	2.4426	0.0146
Years of Q Comp participation	0.0019	0.0015	1.2871	0.1981
Q Comp Levy	-0.0031	0.0012	-2.5087	0.0121
Habitation Minimum distance	0.0024	0.0063	0.3809	0.7033
Q Comp * Habitation Minimum distance	-0.0614	0.0321	-1.9124	0.0558
Distance censored	-0.0052	0.0084	-0.6176	0.5369
Grade 3	0.0055	0.0075	0.7229	0.4698
Grade 5	0.0027	0.0073	0.3692	0.7120
Grade 10	0.0161	0.0070	2.2969	0.0216
English learners	-0.0172	0.0034	-5.0134	0.0000
English learners ^ 2	-0.0143	0.0016	-9.0004	0.0000
Mobility	-0.0569	0.0032	-17.7823	0.0000
Mobility ^ 2	-0.0468	0.0035	-13.5000	0.0000
Mobility ^ 3	0.0071	0.0005	13.1841	0.0000
Poverty	-0.0488	0.0021	-23.4141	0.0000
Poverty ^ 2	-0.0125	0.0008	-16.3139	0.0000
Segregation	-0.0509	0.0030	-17.0926	0.0000
Segregation ^ 2	-0.0022	0.0006	-3.6138	0.0003
Special education	-0.0372	0.0020	-18.1930	0.0000
Special education ^ 2	-0.0418	0.0033	-12.5805	0.0000
Special education ^ 3	0.0039	0.0006	6.4099	0.0000

Table 4.4.2. Reading results: Random effects

Level	Random effect	Variance	Proportion
District	Intercept	0.0159	0.2203
School	Intercept	0.0300	0.4164
Time	Residual	0.0262	0.3633

Q Comp has positively and significantly impacted reading achievement. Schools in Q Comp districts that levied the average amount exhibited 0.0247 standard deviation greater reading achievement on average. There is not enough evidence to conclude that the effect changed over time. Q Comp revenue is significantly negatively associated with

reading achievement, which decreases by -0.0031 standard deviation per allowable levy share (log odds) on average all else being equal.

Sensitivity analyses

Two analyses of sensitivity were undertaken to check the robustness of the overall conclusions about Q Comp. The first analysis mirrors the approach taken by Schwartz (2012) and Sojourner et al. (in press). Instead of treating schools and districts as random intercepts, a weighted least squares model of math and reading scores was estimated with schools as fixed effects (i.e., as J - 1 dummy variables), followed by calculating standard errors that are robust to clustering of schools in districts. When schools are treated as fixed effects, conclusions about distance to the Q Comp border and the program's influence on math and reading achievement do not change (i.e., they are not sensitive), and the parameter estimates are comparable. Appendices 2 and 3 show the results.

The second sensitivity analysis applies a nonequivalent dependent variable approach. Given that Q Comp is supposed to influence student achievement, the treatment should exhibit a positive effect on achievement but no effect on a different student characteristic that Q Comp is not supposed to influence. That is, the results should diverge. If Q Comp influences the nonequivalent dependent variable more than the intended outcome or in unexpected directions, then that would warrant questioning the validity of findings. Segregation was chosen as the nonequivalent dependent variable because it is less endogenous (i.e., unlikely to be manipulated for reasons related to Q Comp) when compared to special education and English-learner status, for example. Additionally, pre-test analyses revealed that segregation was balanced but varied more with distance to the Q Comp border than the other student characteristics, making it a

good case for further exploration. The fixed effects (See Appendices 4.1 and 4.2) explain about 64% of variation in segregation. For a given observation, about 57% of unexplained variation is attributable to the district and about 30% to the school. The results reveal that Q Comp did not significantly influence segregation at the Q Comp border. Within Q Comp districts, years of participation and levy amount are positively associated with segregation, and distance is negatively associated. These findings suggest that even though Q Comp did not have a main effect on segregation, it did vary over time within Q Comp districts. That Q Comp had a significant expected effect on achievement but not a GLATE on segregation lends credibility to the inference about Q Comp's effectiveness.

Value added by Q Comp districts and schools

Tables 4.5.1 and 4.5.2 show results from the math value-added model of correctly allocated Q Comp schools. Because the fixed effects in Equation 4.1 have already explained variation in student achievement, the fixed intercept for residualized math achievement represents the pre-Q Comp intercept of participating schools. It is significantly greater than zero, suggesting that schools with higher-achieving students were more likely to participate in Q Comp. The other fixed effects are constrained to zero. (As a check, Q Comp and years of participation fixed effects were added to the math value-added model found to be insignificant.) Variability in initial math achievement is similar at the district and school levels, but variability in initial and slope interruptions at the school level is greater than at the district level. About 53% of variance in initial status, 32% in initial interruptions and 27% in slope interruptions lie between districts. Negative correlations between initial and slope interruptions at both

district and school levels suggest that variation in math achievement decreases the longer a district and school participates in Q Comp.

Table 4.5.1. Math value-added results: Fixed effect

Fixed effect	Estimate	Standard error	t	p
Intercept	0.0521	0.0241	2.1587	0.0309

Table 4.5.2. Math value-added results: Random effects

Level	Random effect	Covariate	Variance (covariance)	Standard deviation (correlation)
School	Intercept		0.0320	0.1788
School	Q Comp		0.0052	0.0718
School	Years of Q Comp participation		0.0006	0.0238
School	Intercept	Q Comp	(0.0045)	(0.3502)
School	Intercept	Years of Q Comp participation	(-0.0015)	(-0.3478)
School	Q Comp	Years of Q Comp participation	(-0.0007)	(-0.4159)
District	Intercept		0.0356	0.1886
District	Q Comp		0.0024	0.0489
District	Years of Q Comp participation		0.0002	0.0146
District	Intercept	Q Comp	(-0.0008)	(-0.0825)
District	Intercept	Years of Q Comp participation	(-0.0010)	(-0.3500)
District	Q Comp	Years of Q Comp participation	(-0.0002)	(-0.3463)
Time			0.0301	0.1735

As shown in the following tables and figures, three Q Comp districts and five schools exhibited significant math achievement gains. All of the districts and all but one of the schools fall in the second interrupted-time-series category: insignificant initial bump but significant increases in achievement over time.

Table 4.6. Exemplary Q Comp districts: Math

District	Intercept	Q Comp	Years of participation
FARMINGTON PUBLIC SCHOOL DISTRICT	-0.022	-0.038	0.019*
NORTH ST PAUL-MAPLEWOOD SCHOOL DIST	-0.033	0.044	0.016*
ST. FRANCIS PUBLIC SCHOOL DISTRICT	-0.149*	-0.011	0.022*

Figure 4.3. Exemplary Q Comp districts: Math

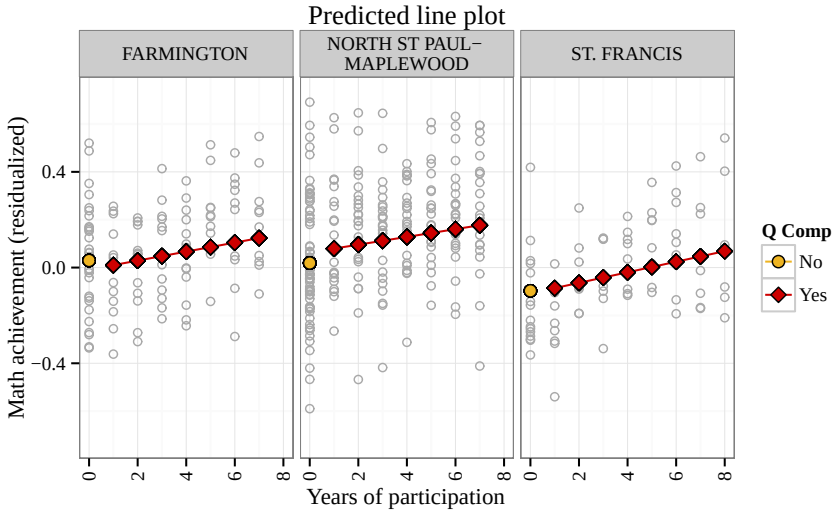
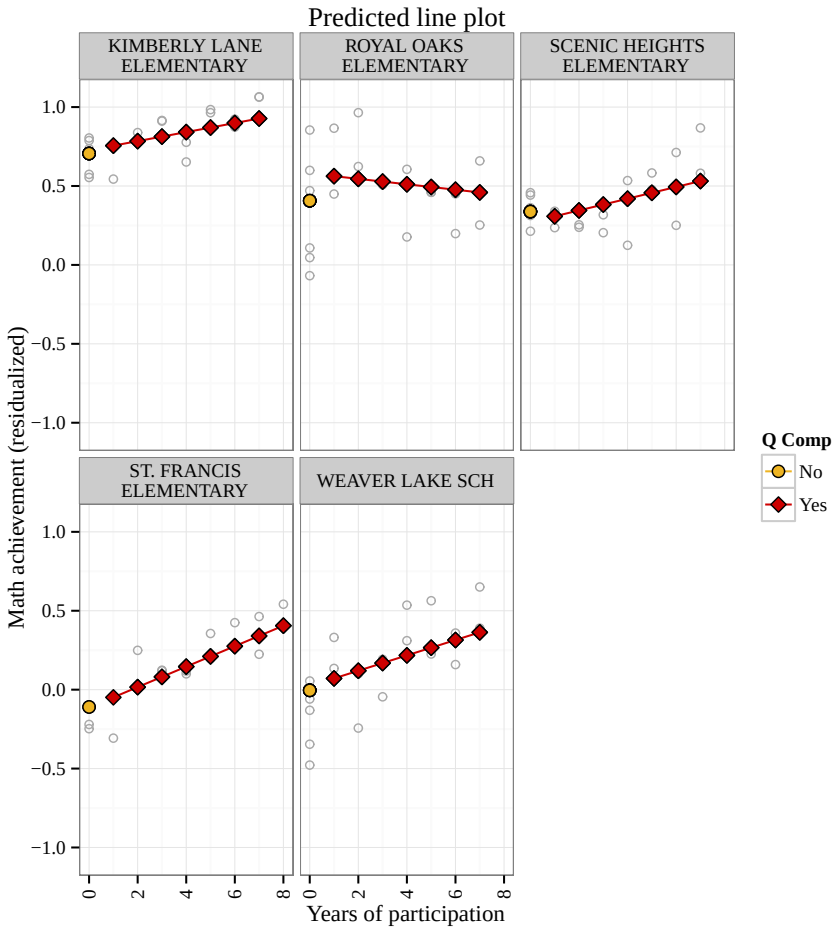


Table 4.7. Exemplary Q Comp schools: Math

District	School	Intercept	Q Comp	Years of participation
MINNETONKA PUBLIC SCHOOL DISTRICT	SCENIC HEIGHTS ELEMENTARY	-0.064	-0.051	0.033*
OSSEO PUBLIC SCHOOL DISTRICT	WEAVER LAKE SCIENCE MATH & TECH SCH	-0.055	0.029	0.053*
SOUTH WASHINGTON COUNTY SCHOOL DIST	ROYAL OAKS ELEMENTARY	0.323*	0.146*	-0.027
ST. FRANCIS PUBLIC SCHOOL DISTRICT	ST. FRANCIS ELEMENTARY	-0.012	0.007	0.043*
WAYZATA PUBLIC SCHOOL DISTRICT	KIMBERLY LANE ELEMENTARY	0.176*	0.027	0.035*

Figure 4.4. Exemplary Q Comp schools: Math



Tables 4.5.1 and 4.5.2 show results from the reading value-added model of correctly allocated Q Comp schools. Because the fixed effects in Equation 4.1 have already explained variation in student achievement, the fixed intercept for residualized reading achievement represents the pre-Q Comp intercept of participating schools. It is significantly greater than zero, suggesting that schools with higher-achieving students were more likely to participate in Q Comp. The other fixed effects are constrained to zero. (As a check, Q Comp and years of participation fixed effects were added to the reading value-added model found to be insignificant.) Variability in initial reading

achievement is similar at the district and school levels, but variability in initial and slope interruptions at the school level is greater than at the district level. About 47% of variance in initial status, 22% in initial interruptions and 10% in slope interruptions lie between districts. After entering Q Comp, variability in reading achievement decreases over time at the school level and while increasing slightly at the district level, as suggested by the correlations between initial and slope interruptions.

Table 4.8.1. Reading value-added results: Fixed effect

Fixed effect	Estimate	Standard error	t	p
Intercept	0.0455	0.0208	2.1868	0.0288

Table 4.8.2. Reading value-added results: Random effects

Level	Random effect	Covariate	Variance (covariance)	Standard deviation (correlation)
School	Intercept		0.0265	0.1627
School	Q Comp		0.0045	0.0670
School	Years of Q Comp participation		0.0003	0.0171
School	Intercept	Q Comp	(-0.0002)	(-0.0170)
School	Intercept	Years of Q Comp participation	(-0.0011)	(-0.4035)
School	Q Comp	Years of Q Comp participation	(-0.0004)	(-0.3510)
District	Intercept		0.0237	0.1539
District	Q Comp		0.0013	0.0360
District	Years of Q Comp participation		0.0000	0.0057
District	Intercept	Q Comp	(-0.0009)	(-0.1600)
District	Intercept	Years of Q Comp participation	(0.0002)	(0.1932)
District	Q Comp	Years of Q Comp participation	(0.0000)	(0.1420)
Time			0.0209	0.1446

As shown in the following tables and figures, two Q Comp districts and six schools exhibited significant reading achievement gains. Both districts and one school

exhibited an initial increase that subsequently remained steady over time. The remaining schools did not exhibit an initial bump, but reading achievement increased significantly over time.

Table 4.9. Exemplary Q Comp districts: Reading

District	Intercept	Q Comp	Years of participation
OSSEO PUBLIC SCHOOL DISTRICT	-0.025	0.039*	0.000
SPRING LAKE PARK PUBLIC SCHOOLS	-0.071	0.053*	0.002

Figure 4.5. Exemplary Q Comp districts: Reading

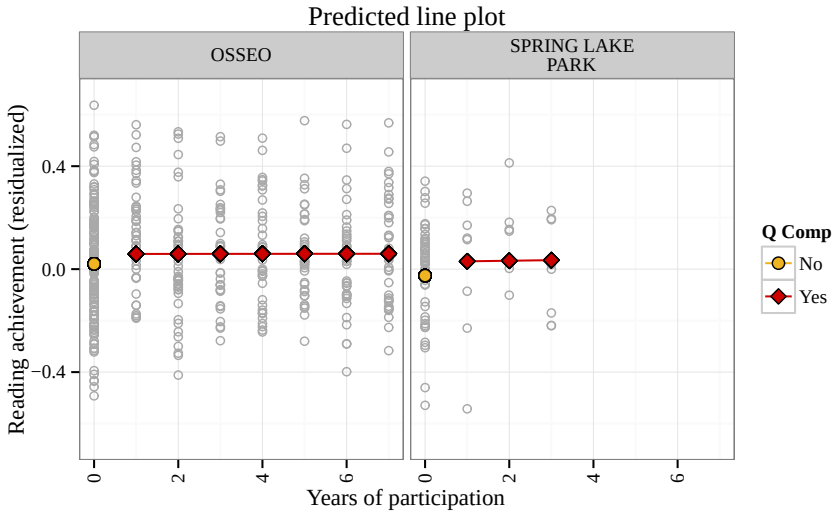
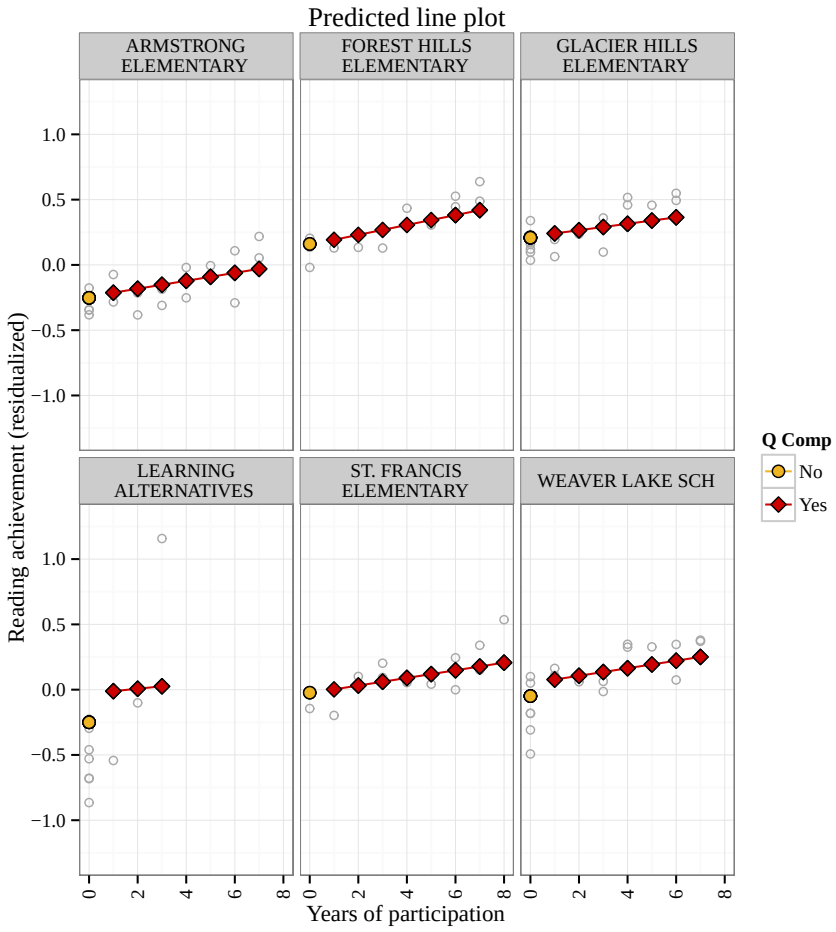


Table 4.10. Exemplary Q Comp schools: Reading

District	School	Intercept	Q Comp	Years of participation
EDEN PRAIRIE PUBLIC SCHOOL DISTRICT	FOREST HILLS ELEMENTARY	-0.118	-0.002	0.031*
OSSEO PUBLIC SCHOOL DISTRICT	WEAVER LAKE SCIENCE MATH & TECH SCH	-0.070	0.059	0.029*
ROSEMOUNT-APPLE VALLEY-EAGAN	GLACIER HILLS ELEMENTARY	-0.038	0.018	0.029*
SOUTH WASHINGTON COUNTY SCHOOL DIST	ARMSTRONG ELEMENTARY	-0.322*	-0.022	0.026*
SPRING LAKE PARK PUBLIC SCHOOLS	LEARNING ALTERNATIVES COMMUNITY SCH	-0.225*	0.166*	0.016
ST. FRANCIS PUBLIC SCHOOL DISTRICT	ST. FRANCIS ELEMENTARY	0.013	-0.001	0.030*

Figure 4.6. Exemplary Q Comp schools: Reading



Schwartz (2012) interviewed key informants at two school districts that added significant value to student achievement. Farmington emphasized *peer teachers* as the source of new instructional knowledge, and St. Francis emphasized *external sources of new knowledge*. Those strategies and the fidelity with which they were applied may help explain why those districts were able to add value, but more research is needed to identify the strategies used by exemplary districts and schools that accelerated student achievement.

Chapter 5: Summary and Conclusions

Estimating causal effects is an important aim when program evaluation questions pertain to a program's measurable objectives. Randomized field experiments are considered a gold standard for internal (and external) validity (Boruch, 1991). However, rather than randomly assigning participants to a treatment or control group, many programs are implemented in geographically defined jurisdictions, such as school districts in the case of Minnesota's Quality Compensation for Teachers (Q Comp) program. Some districts have chosen to participate in Q Comp, but most have not. How have students in Q Comp districts and schools fared relative to students in sites that did not reform teacher pay and professional development? It is an important educational policy question, but is it appropriate to evaluate Q Comp by simply comparing participants to nonparticipants?

Following Holmes (1998) and others, this study has attempted to distinguish Q Comp's impact on student achievement from other influences by limiting the comparison to the geographic border separating participating schools from neighboring schools that did not participate. In doing so, it has addressed two sets of evaluation questions.

1. Has Q Comp been effective? To what degree has Q Comp led to student achievement gains as theorized? Which Q Comp districts and schools added significant value and warrant emulation?
2. Is applying spatial regression discontinuity (SRD) worthwhile (given the effort and costs to parsimony and external validity)? To what degree does SRD yield well-matched comparisons (i.e., counterfactuals) at the Q Comp border on

average? To what degree does the SRD approach, with design and analytic enhancements, rule out validity threats?

It should be stressed that although this dissertation describes SRD, integrates related theories, and demonstrates best practices, it does not present a new methodology as much as contribute to evaluation practice, especially as it relates to evaluating Q Comp and other educational policies implemented by a subset of school districts within a state.

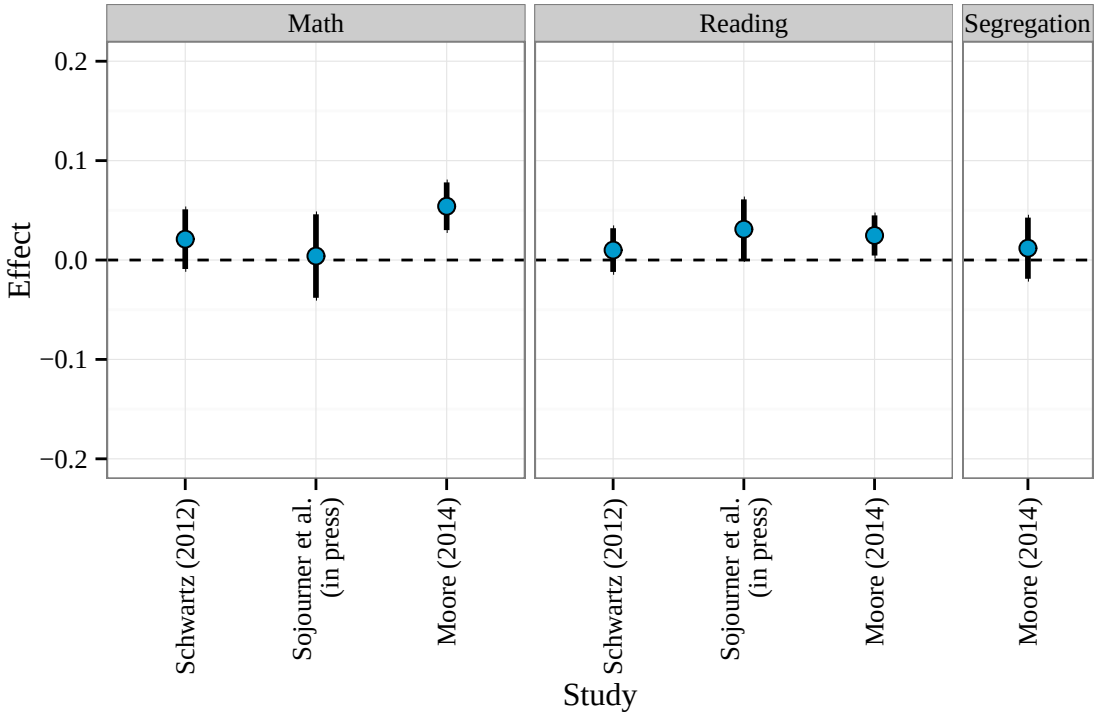
Q Comp

Q Comp has successfully overcome a number of difficulties to implement enduring changes to teacher pay and professional development but in ways that are difficult to characterize and measure. The program could have faltered under the weight of competing political demands, especially given MDE's lack of funding and capacity, but program documentation, surveys and interviews indicate that Q Comp has struck a balance between performance pay advocated by incentive-focused designers and job-embedded professional development favored by teachers. And it has managed to attract a growing number of districts without surpassing budget allocations.

Based on the SRD models specified, Q Comp had a significantly positive main effect on math and reading achievement (see Figure 5.1). Participating in Q Comp is associated with an increase of 0.0541 standard deviation (*se* = 0.012) in math achievement and an increase of 0.0247 standard deviation (*se* = 0.01) in reading achievement on average all else being equal. Following Sojourner et al.'s (in press) cost-benefit analysis of reading achievement, Q Comp's social benefit is estimated to outweigh its cost by roughly 4-to-1 (i.e., $24,700 / $6,500). Q Comp's impact is neither geographically local nor time-varying, but levy share is a significantly negative

moderator. The findings are not sensitive to treating schools as fixed effects, and they diverge from a nonequivalent dependent variable (i.e., segregation).

Figure 5.1. Estimates of Q Comp's impact (with confidence intervals) on student achievement and a non-equivalent dependent variable



Earlier studies of Q Comp did not find a significant math impact, but one found an impact on reading that was similar in magnitude to this study. Why might conclusions in this study differ from findings by Schwartz (2012) and Sojourner et al. (in press)? Even though all three studies longitudinally analyze state-mandated test scores from pre- and post-test years, this study's sample and methods differ in key ways. The other studies' samples did not include grades 10 or 11 or outcomes from recent years (i.e., 2011-2013), and they did not limit the sample to Q Comp districts and contiguous neighbors. Sojourner et al. (in press) had access to student test scores. This study applies

different weights, mixed-effects modeling instead of least-squares and transformations of proportions. Lastly, the other studies did not estimate the effect of Q Comp levy share.

Table 5.1. Study characteristics that may account for differences in Q Comp estimates

Characteristic	Schwartz (2012)	Sojourner et al. (in press)	Moore (2014)
Grade levels	3, 5, and 7	3, 5, and 7	3, 5, 7, 10 (reading), and 11 (math)
School years	2004-2010	2004-2010	2004-2013
Comparison group	All non-participating school districts	All non-participating school districts	Neighboring non-participating school districts
MCA scores	School means	Student scores	School means
Modeling	Student-weighted least squares	Ordinary least squares	Mixed-effects weighted by square root of students
Transformed proportions	No	No	Yes
Levy share	No	No	Yes

Which Q Comp districts and schools added significant value and warrant emulation? Five school districts exhibited student achievement gains that coincided with Q Comp participation: Farmington, North St. Paul-Maplewood, Osseo, Spring Lake Park, and St. Francis. The schools that added significant value were Armstrong Elementary (South Washington County district), Forest Hills Elementary (Eden Prairie), Glacier Hills Elementary (Rosemount-Apple Valley-Eagan), Kimberly Lane Elementary (Wayzata), Learning Alternatives Community School (Spring Lake Park), Royal Oaks Elementary (South Washington County), Scenic Heights Elementary (Minnetonka), St. Francis Elementary (St. Francis), and Weaver Lake Science Math & Tech School (Osseo).

Spatial regression discontinuity

Although Q Comp qualifies as a naturally designed spatial regression discontinuity (SRD) study whereby distance to a geographic border essentially determines program participation, regressing student achievement on piecewise distance

accomplished almost nothing over the other quasi-experimental approaches employed. That is, a simpler analysis of covariance suffices for explaining how student achievement varied over time with program participation while accounting for competing explanations (e.g., socioeconomic characteristics) not ruled out by design. In addition to statistical controls and longitudinal data, other quasi-experimental enhancements that proved valuable include 1) using pre-test data to choose an appropriate definition and functional form for distance, 2) using pre-test data to establish that student achievement was ignorably different at the border before Q Comp, and 3) excluding highly distal observations from the comparison group. Failing to reject the null hypotheses about the influence of distance had at least one benefit: it revealed that Q Comp's average treatment effects are not geographically local (i.e., they generalize to schools beyond the border cutoff, although not necessarily beyond Q Comp districts and their contiguous neighbors).

When is it worthwhile to design and analyze spatial regression discontinuity studies? Future research is needed to answer that question, but as discussed earlier, pre-test data are highly valuable for prospective design and choosing a functional form. Pre-test outcomes and distance can be used to choose a cutoff that enhances statistical power, prioritizes treatment to those most in need, and/or simplifies distance calculations (i.e., by drawing straight cutoffs). In the case of natural experiments, the decision to apply a SRD analysis is more likely to hinge on the fixed cost of operationalizing and quantifying distance because, if it is low, then it follows that the overall cost of pre-test exploration and including piecewise distance in the final model is also likely to be low. The benefits of applying spatial regression discontinuity could depend on what is being studied and/or

the selection mechanism. For example, the pre-test analyses of student characteristics and the studies by Holmes (1998), Black (1999) and Bayer, et al. (2007) reveal that treating distance and borders as assignment variables and cutoffs, respectively, can help explain student movements between schools and/or geographic arbitrage.

Limitations and future directions

Several limitations persist and temper the internal and external validity of conclusions about Q Comp. Neither Q Comp adoption nor the SRD approach address the history validity threat. That is, districts and schools were free to implement Q Comp *and* other reforms simultaneously, and the other reforms could be responsible for the observed effect on student achievement. Fuzzy RD and measurement error also threaten the validity of findings. That is, misallocation and unobserved aspects of implementation, especially professional development particulars, are probably attenuating the estimated effects. Another measurement limitation is the use of school means to evaluate Q Comp's impact on student achievement because students' scores were not available.

The objectives-oriented and quantitative approaches also have their limitations. Objectives-oriented evaluation has been criticized for oversimplifying details and emphasizing outcomes over needs and processes. It was not feasible to apply evaluation approaches from the intuition-pluralist end of the evaluation continuum or to collect and analyze qualitative data. As demonstrated by Schwartz (2012), much can be learned about Q Comp via semi-structured interviews of key informants, but this study only used available quantitative data. Utility is another limitation of this study. Local Q Comp stakeholders and researchers within and beyond Minnesota are the intended audiences,

but this study prioritized validity and accuracy over identifying and engaging a wide range of stakeholders to maximize its usefulness.

There are several opportunities for future research and evaluation. The most pressing among them is to apply alternative approaches and methods to the question, "Which components, combinations, and intensity of components (observations, professional learning communities, career advancement, student learning goal setting, performance pay, and/or alternative salary schedules) were most effective at accelerating student achievement among Q Comp participants?" And more generally, "Which components should be prioritized to accelerate student achievement going forward and in other settings?" Even though this evaluation infers that *participation* in Q Comp has been effective overall, the effects are small. And it remains unclear which aspects of Q Comp could be prioritized to magnify its effect. Participation does not guarantee authentic reform of incentives or instructional supports. Conversely, districts can effectively apply and align incentives with instructional supports without participating in Q Comp. Indeed, some districts were already applying Q Comp components before the additional revenue became available, and some have since applied Q Comp strategies without participating in the program. Interviews are labor intensive, but results of the value-added analysis make it more feasible to qualitatively illuminate promising practices by interviewing managers, staff, and other stakeholders at Q Comp sites that accelerated student achievement. Do their practices differ from other participants? Do those differences help explain their success? Given that more residual variation was explained at the district level than at the school level, district interviews might be prioritized over school interviews. Schwartz (2012) has already conducted interviews at two of the

districts that were later identified as exemplary; that leaves three exemplary districts where additional qualitative inquiry could be directed.

Another line of future research involves identifying promising practices that do not already exist within Q Comp. For example, what would happen if MDE took steps to align district interests with its interests by paying Q Comp districts for performance rather than guaranteeing a fixed amount of state aid? Participating districts receive a funding incentive that varies with the number of students and levy approvals (in different ratios over time), but not with district-level performance. Paying districts for performance could have positive effects and/or unintended consequences, such as discouraging district participation. What about job redesign (e.g., "flipped" classrooms) instead of or in addition to professional development? Performance is difficult to operationalize and measure, let alone use as a basis for pay. A new teacher evaluation law in Minnesota is requiring evaluations for the first time starting this year. Future research could examine the role and scientific merits of the new teacher evaluation requirements as a moderator of Q Comp's effectiveness. For example, the new requirements may allow districts to move further away from paying for effort/inputs and closer to paying for actual performance/outputs. Sojourner et al. (in press) stress the importance of finding how Q Comp and other P4P-centered TIS programs influence recruiting, retaining and separating individual employees in order to move the needle of average effectiveness. Teacher evaluation training (i.e., of observers), measuring teacher effectiveness, and data reporting (to teachers and administrators) are other mechanisms that should be explored in greater detail.

Another opportunity would be to update this study as long as Q Comp continues and new school years pass, which is one of the benefits of using available data published by MDE. An updated study could examine the role of levy share in greater detail, given that it is negatively correlated with student achievement. Additionally, it would be interesting to specify and compare grade level models (instead of controlling for grade levels). Charter schools that have participated in Q Comp could be included. Did participating charter schools impact student achievement under Q Comp more than public schools? Which charter schools added value and may be worthy of emulation? Comparing district/school effect sizes to charter school effects may further illuminate human resource strategies that accelerate student achievement (e.g., strategies that charter schools have successfully innovated). After applying qualitative evaluation methods and enhancing the current study, consumer- and expertise-oriented evaluation approaches could then be applied to help MDE refine its interpretation of Q Comp's requirements. That is, after better identifying practices that accelerate student achievement and with the involvement of experts from P4P-centered TIS programs in other states, MDE could knowledgeably revise its application and compliance criteria (i.e., its checklist for district consumers).

Bibliography

Alvarez, L. (2011, Mar 9). In Florida, push to link teacher pay to student performance. *New York Times,* pp. A.19. Retrieved from http://www.nytimes.com/2011/03/09/education/09florida.html

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics, 26*(1), 77-104.

Baddeley, A., & Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software 12*(6), 1-42.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: Chapman & Hall/CRC.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R Foundation for Statistical Computing, Vienna, Austria.

Bayer, P., Ferreira, F., & McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy, 115*(4), 588-638.

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. New York: Springer.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics, 114*(2), 577-599.

Boruch, R. F. (1991). The president's mandate: Discovering what works and what works better. In M. W. McLaughlin & D. C. Phillips (Eds.), *Ninetieth yearbook of the*

National Society for the Study of Education, Part II. Evaluation and education: At quarter century (pp. 147-167). Chicago, IL: The University of Chicago Press.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Brown, T. L. (2005). Evaluating geographic program performance analysis. *Public Performance & Management Review, 29*(2), 164-190.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*(4), 409-429.

Cappelleri, J. C., Darlington, R. B., Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review, 18*, 141–152.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review, 84*(4), 772-793.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review, 14*(1), 5-26.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly journal of economics, 126*(4), 1593-1660.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cook, T. D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Ninetieth yearbook of the National Society for the Study of Education, Part II. Evaluation*

and education: At quarter century (pp. 115-144). Chicago, IL: The National Society for the Study of Education.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142*(2), 636-654.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines*. Boston: Pearson.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1), 1-53.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211.

Goldberger, A. S. 1972. *Selection bias in evaluating treatment effects: Some formal illustrations*. Discussion paper number 123. Madison, WI: Institute for Research on Poverty.

Graham, E. T., (2001). *A methodological framework for neighborhood indicators: Using spatial probability techniques and multi-attribute utility analysis to evaluate the*

effectiveness of a neighborhood-based child maltreatment prevention program. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Hahn, J., Todd, P., Van Der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica 69*(1), 201–209.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal, 113*(485), F64-F98.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466-479.

Herrmann, M., Walsh, E., Isenberg, E., Resch, A. (2013). *Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/publications/PDFs/education/value-added_shrinkage_wp.pdf

Hezel Associates. (2009). *Quality Compensation for Teachers: Summative evaluation*. Saint Paul, MN: Minnesota Department of Education. Retrieved from http://education.state.mn.us/mdeprod/groups/Communications/documents/Report/036790.pdf

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Holmes, T. J. (1998). The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy, 106*(4), 667-705.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Institute of Education Sciences. (2007). *Cluster randomized trials: Presentations*. Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncer/whatsnew/conferences/rct_traininginstitute/presentations.asp

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, 267-292.

Keele, L., & Titiunik, R. (2011). *Geographic boundaries as regression discontinuities*. Retrieved from http://www.personal.psu.edu/ljk20/GeoRDD.pdf

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal, 113*(485), F34-F63.

Laws of Minnesota First Special Session 2005, chapter 5, art. 2, sec. 40. Retrieved from https://www.revisor.mn.gov/laws/?id=5&doctype=Chapter&year=2005&type=1

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*, 281-355.

Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin, 126*(2), 309-337.

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Thousand Oaks, Calif: SAGE.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698.

Minnesota Department of Education. (2014). *Data Center.* Saint Paul, MN: Author. Retrieved from http://education.state.mn.us/MDE/Data/index.html

Minnesota Geospatial Information Office. (2013). *MnGeo's Clearinghouse Data Catalog*. Saint Paul, MN: Author. Retrieved from http://www.mngeo.state.mn.us/chouse/metalong.html

Minnesota Statutes 122A.413-122A.415. (2009). Retrieved from https://www.revisor.mn.gov/statutes/?id=122A.413

Monmonier, M. (1996). *How to lie with maps*. Chicago: University Of Chicago Press.

Moore, C. T. (2009). *Spatial regression discontinuity: Estimating effects of geographically implemented programs and policies*. Presented at the annual conference of the American Evaluation Association, Orlando, FL.

Office of the Legislative Auditor. (2009). *Q Comp: Quality Compensation for Teachers*. Saint Paul, MN: Author. Retrieved from http://www.auditor.leg.state.mn.us/ped/2009/Q Comp.htm

Orfield, M. & Wallace, N. (2007). Expanding educational opportunity through school and housing choice. *CURA Reporter, 37*(2), 19-26. Retrieved from: http://www.cura.umn.edu/reporter/07-Summ/Summ-07-Issue.pdf

Ormsby, T., Napoleon, E., Burke, R., & Groessl, C. (2001). *Getting to know ArcGIS Desktop: The basics of ArcView, ArcEditor, and ArcInfo*. Redlands, CA: ESRI Press.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics, 161*(2), 203-207.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text*. Thousand Oaks, CA: Sage Publications.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909-950.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37*(1), 7-63.

Quon Huber, M. S., Van Egeren, L. A., Pierce, S. J., & Foster-Fishman, P. G. (2009). GIS applications for community-based research and action: mapping change in a community-building initiative. *Journal of prevention & intervention in the community, 37*(1), 5-20.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renger, R., Cimetta, A., Pettygrove, S., & Rogan, S. (2002). Geographic information systems (GIS) as an evaluation tool. *American Journal of Evaluation, 23*(4), 469.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688.

Schwartz, N. L. (2012). *Aligning teacher improvement strategies: A mixed method study of teaching reform in Minnesota*. (Doctoral dissertation). Ann Arbor, MI: University of Michigan.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Sojourner, A., Mykerezi, E., & West, K. (in press). Teacher pay reform and productivity: Panel data evidence from adoptions of Q-Comp in Minnesota. *Journal of Human*

Resources. Retrieved from https://sites.google.com/site/aaronsojourner/Sojourner%20Mykerezi%20West%20-%20QComp.pdf?attredirects=0

Talen, E., & Shah, S. (2007). Neighborhood evaluation using GIS: An exploratory study. *Environment and Behavior, 39*(5), 583-615.

Tate, William F. (2008). Geography of opportunity: Poverty, place, and educational outcomes. *Educational Researcher, 37*(7), 397-411.

U.S. Census Bureau. (2012). Technical Documentation: 2010 TIGER/Line Shapefiles. Washington, DC: Author. Retrieved from http://www.census.gov/geo/maps-data/data/tiger-line.html

U.S. Department of Education. (2005). Scientifically based evaluation methods: Notice of final priority. *Federal Register, 70*(15), 3585-3589. Retrieved from http://www.ed.gov/legislation/FedRegister/finrule/2005-1/012505a.html

U.S. Department of Education. (2011). *Race to the Top Fund*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *The American Economic Review, 99*(1), 179-215.

Value-Added Research Center. (2013). *Technical report: Minneapolis value-added model*. Madison, WI: Author.

Verdi, M. P., & Kulhavy, R. W. (2002). Learning with maps and texts: An overview. *Educational Psychology Review, 14*(1), 27-46.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. Hoboken, NJ: Wiley.

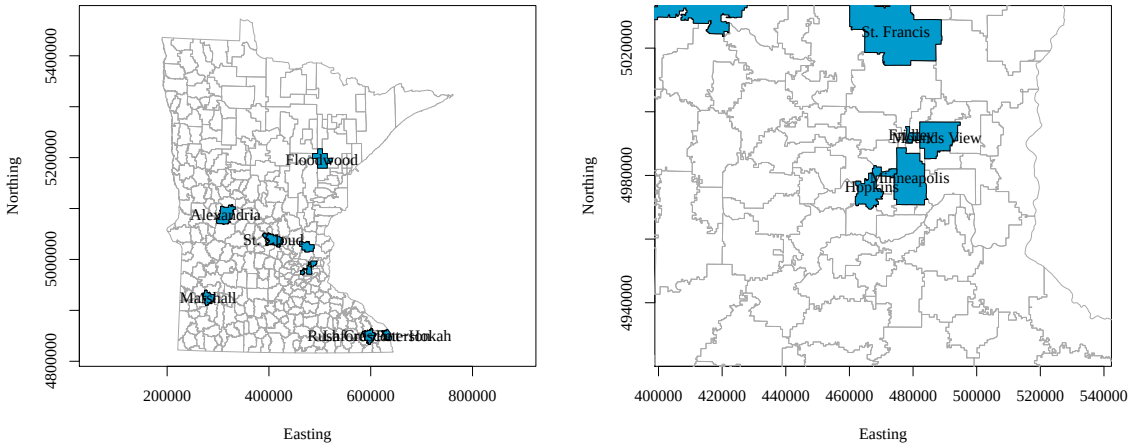
Wong, V. C., Steiner, P. M., & Cook, T. D. (2010). *Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods*. Evanston, IL: Northwestern University. Retrieved from http://www.ipr.northwestern.edu/publications/papers/WongSteinerCook-MRDD.pdf

Yarbrough, D. B., & Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users*. Thousand Oaks, Calif: SAGE.

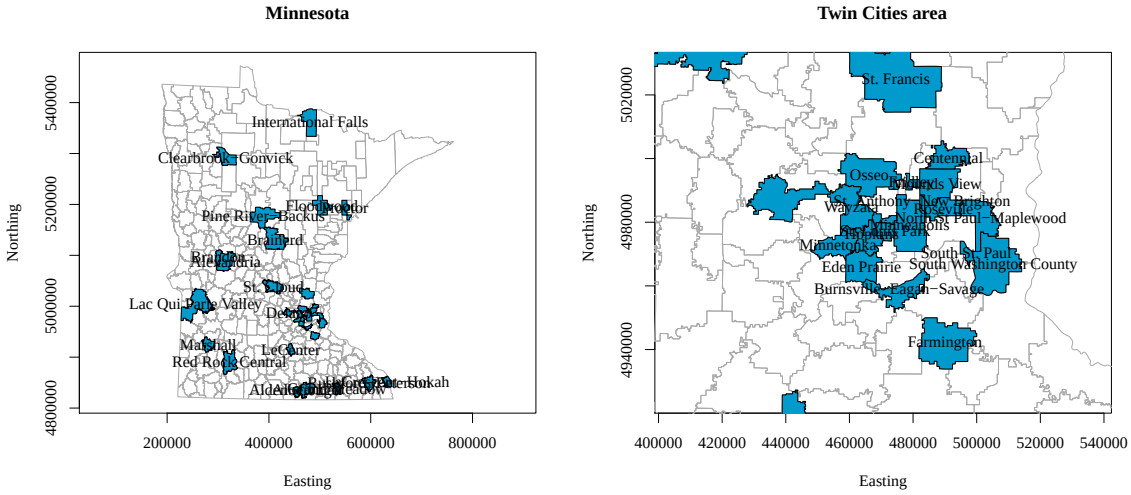
Appendices

Appendix 1. Maps of Q Comp districts by year

2006

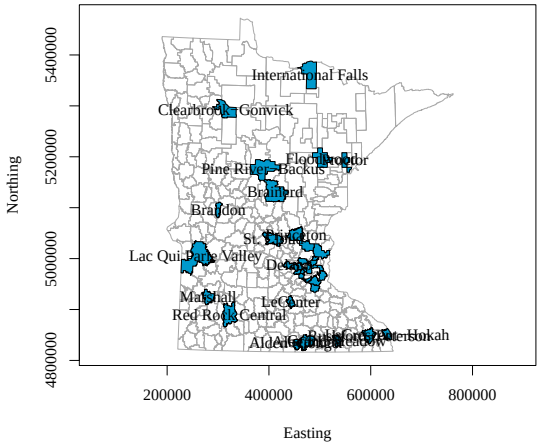


2007

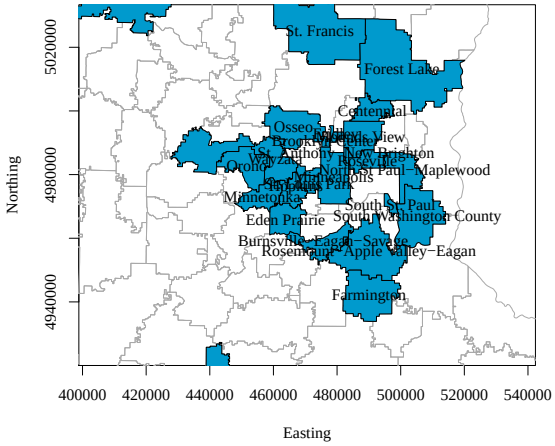


2008

Minnesota

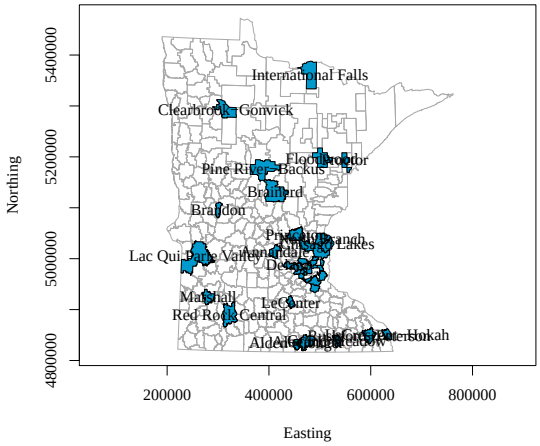


Twin Cities area

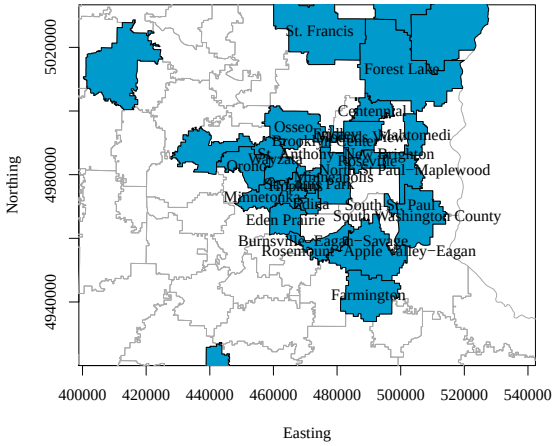


2009

Minnesota

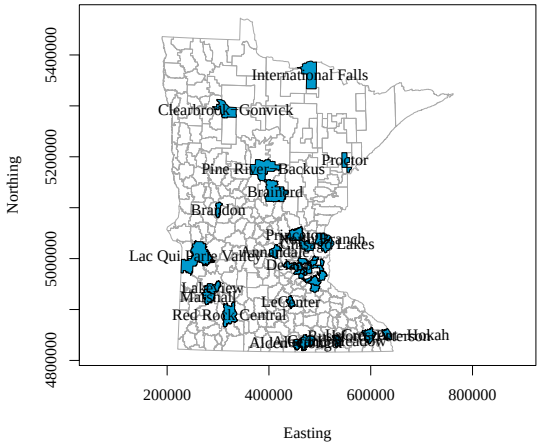


Twin Cities area

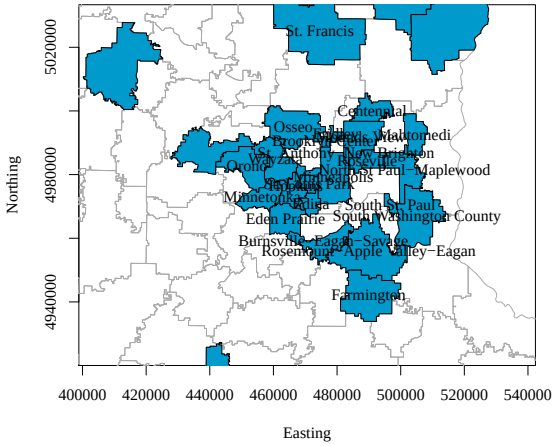


2010

Minnesota

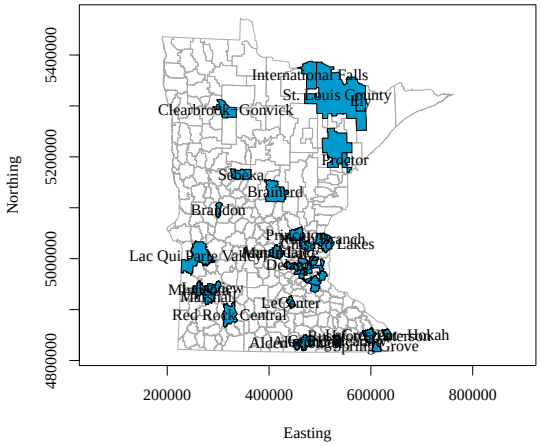


Twin Cities area

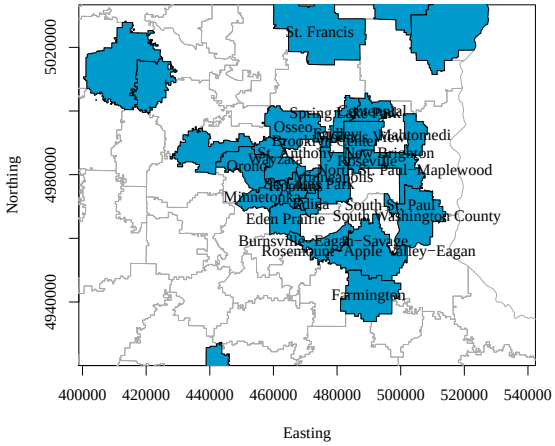


2011

Minnesota

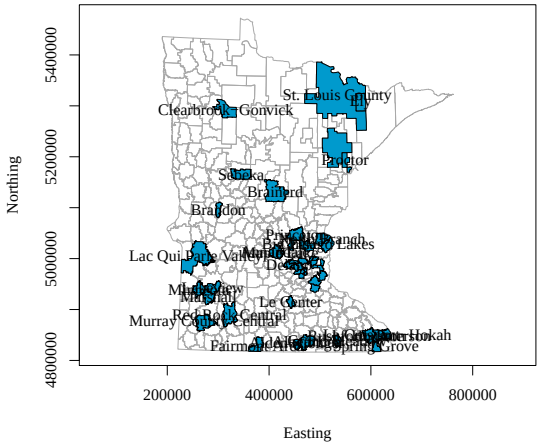


Twin Cities area

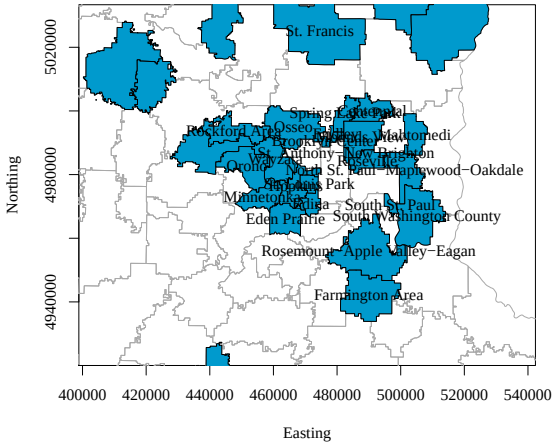


2012

Minnesota

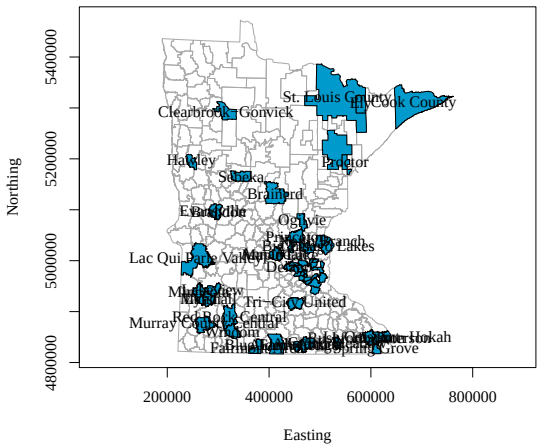


Twin Cities area

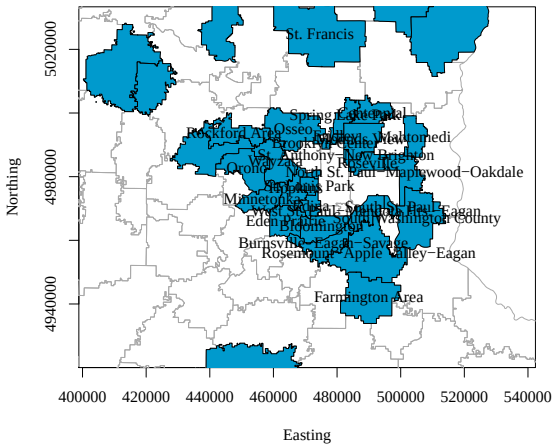


2013

Minnesota



Twin Cities area



Appendix 2. Math weighted least squares results

	Estimate	Standard error	t	p
Intercept	-0.0416	0.0264	-1.5731	0.1157
Year centered on 2004	0.0015	0.0007	2.1278	0.0334
Q Comp	0.0513	0.0116	4.4139	0.0000
Years of Q Comp participation	-0.0020	0.0016	-1.2092	0.2266
Q Comp Levy	-0.0074	0.0014	-5.3934	0.0000
Habitation Minimum distance	-0.0061	0.0071	-0.8518	0.3943
Q Comp * Habitation Minimum distance	-0.0368	0.0367	-1.0024	0.3161
Distance censored	-0.0062	0.0093	-0.6672	0.5046
Grade 3	-0.0768	0.0109	-7.0588	0.0000
Grade 5	-0.0743	0.0104	-7.1171	0.0000
Grade 11	0.0416	0.0087	4.8018	0.0000
English learners	-0.0071	0.0043	-1.6674	0.0954
English learners ^ 2	-0.0133	0.0026	-5.1906	0.0000
Mobility	-0.0468	0.0038	-12.2589	0.0000
Mobility ^ 2	-0.0161	0.0042	-3.8525	0.0001
Mobility ^ 3	0.0034	0.0006	5.3321	0.0000
Poverty	-0.0451	0.0028	-15.9606	0.0000
Poverty ^ 2	-0.0121	0.0010	-11.9586	0.0000
Segregation	-0.0419	0.0037	-11.3142	0.0000
Segregation ^ 2	-0.0009	0.0008	-1.2011	0.2297
Special education	-0.0401	0.0026	-15.3148	0.0000
Special education ^ 2	-0.0361	0.0040	-9.0953	0.0000
Special education ^ 3	0.0049	0.0010	4.9669	0.0000

Note: School fixed effects are not shown. $R^2 = 0.823$.

Appendix 3. Reading weighted least squares results

	Estimate	Standard error	t	p
Intercept	-0.0885	0.0273	-3.2469	0.0012
Year centered on 2004	0.0021	0.0006	3.4168	0.0006
Q Comp	0.0255	0.0096	2.6734	0.0075
Years of Q Comp participation	0.0015	0.0013	1.1660	0.2436
Q Comp Levy	-0.0033	0.0012	-2.7539	0.0059
Habitation Minimum distance	-0.0005	0.0060	-0.0825	0.9343
Q Comp * Habitation Minimum distance	-0.0533	0.0305	-1.7485	0.0804
Distance censored	-0.0086	0.0079	-1.0828	0.2789
Grade 3	-0.0754	0.0090	-8.3460	0.0000
Grade 5	-0.0730	0.0087	-8.4313	0.0000
Grade 10	0.0688	0.0076	8.9920	0.0000
English learners	-0.0248	0.0034	-7.2935	0.0000
English learners ^ 2	-0.0081	0.0020	-4.1104	0.0000
Mobility	-0.0408	0.0033	-12.4474	0.0000
Mobility ^ 2	-0.0210	0.0039	-5.3357	0.0000
Mobility ^ 3	0.0039	0.0006	6.5029	0.0000
Poverty	-0.0403	0.0026	-15.7456	0.0000
Poverty ^ 2	-0.0100	0.0009	-11.1208	0.0000
Segregation	-0.0433	0.0032	-13.6220	0.0000
Segregation ^ 2	-0.0005	0.0007	-0.8280	0.4077
Special education	-0.0415	0.0023	-17.7095	0.0000
Special education ^ 2	-0.0348	0.0033	-10.3932	0.0000
Special education ^ 3	0.0051	0.0010	5.1037	0.0000

Note: School fixed effects are not shown. $R^2 = 0.855$.

Appendix 4.1.1. Segregation non-equivalent dependent variable results: Fixed effects

Fixed effect	Estimate	Standard error	t	p
Intercept	-0.8276	0.0749	-11.0503	0.0000
Year centered on 2004	0.0448	0.0016	27.3301	0.0000
Q Comp	0.0216	0.0279	0.7734	0.4393
Years of Q Comp participation	0.0168	0.0040	4.2459	0.0000
Q Comp Levy	0.0116	0.0033	3.4759	0.0005
Habitation Minimum distance	0.0296	0.0170	1.7365	0.0825
Q Comp * Habitation Minimum distance	-0.1754	0.0886	-1.9801	0.0477
Distance censored	0.0147	0.0225	0.6531	0.5137
Grade 3	-0.0288	0.0238	-1.2071	0.2274
Grade 5	-0.0199	0.0231	-0.8635	0.3879
Grade 10	-0.0753	0.0197	-3.8257	0.0001
Grade 11	-0.1211	0.0198	-6.1227	0.0000
English learners	0.1335	0.0096	13.9534	0.0000
English learners ^ 2	0.0181	0.0044	4.0810	0.0000
Mobility	0.1305	0.0087	15.0826	0.0000
Mobility ^ 2	0.0541	0.0088	6.1344	0.0000
Mobility ^ 3	-0.0062	0.0013	-4.6770	0.0000
Poverty	0.1285	0.0055	23.2357	0.0000
Poverty ^ 2	0.0450	0.0020	22.1444	0.0000
Special education	-0.0334	0.0058	-5.7995	0.0000
Special education ^ 2	-0.0306	0.0092	-3.3310	0.0009
Special education ^ 3	0.0077	0.0017	4.4566	0.0000

Appendix 4.1.2. Segregation non-equivalent dependent variable results: Random effects

Level	Random effect	Variance	Proportion
District	Intercept	0.9214	0.5666
School	Intercept	0.4821	0.2964
Time	Residual	0.2228	0.1370