THE ACHIEVEMENT AND NON-ACHIEVEMENT EFFECTS OF REPEATING ANOTHER YEAR WITH A TEACHER AND REVERSING BROKEN WINDOWS THEORY

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
James M. Kelly
December 2017

Examining Committee Members:

Douglas A. Webber, Advisory Chair, Economics
J. Catherine Maclean, Economics
Michael A. Leeds, Economics
Sarah A. Cordes, External Member, Temple University College of Education

ABSTRACT

This research gives a multidimensional investigation into community policies that are becoming more prevalent in American society. In Chapter 1, I apply multiple Value-Added Models (VAM) of achievement to data from the North Carolina Education Research Data Center (NCERDC) to determine the academic impacts of repeating a year (or more) with the same teacher on student achievement in math and reading. Given the growing trend in schools and teaching practices, like looping, that pair teachers and students for multiple years, this research finds contrasting results about the gains in academic achievement associated with repeating with a teacher. Specifically, while there is evidence that students on average have higher scores when repeating with a teacher, this effect is mitigated when one controls for teacher quality. Using limited probability models, I find students are 29\%-34\% and 42\%-46\% more likely to repeat with a teacher whose Value-Added estimate is in the top 20\% of teacher-quality compared to a teacher in the bottom 20\% in math and reading, respectively. This nonrandom assignment of students to teachers, creates upward bias in the estimated achievement effects of repeating with a teacher that have previously been unaccounted for. In chapters 1 and 2, I account for nonrandom assignment finding non-significant gains in *achievement* associated with repeating with a teacher.

While Chapter 1 finds non-significant gains to student achievement, Chapter 2 investigates if there are any non-cognitive gains students experience when they repeat with a teacher for another year. Using the same longitudinal data from the NCERDC, Chapter 2's results indicate increases in character-trait measures associated with teacher and student perceptions of academic success and effort. Using multiple partial persistence VAMs that

include controls for student heterogeneity and for teacher quality, the estimated effects on a teacher's subjective scoring of a student's academic success, student's anticipated grade for the year, and student attendance are all significantly greater than zero. Taken together, the positive effects from students repeating with the same teacher reveal themselves prevalently on character-trait improvements rather than on contemporaneous achievement scores.

In Chapter 3, I investigate the causal direction of a popular policing policy. Although there are a large number of studies testing Broken Windows Theory (BWT) (Wilson & Kelling, 1982), the reverse theoretical pathway has never been assessed (risk perceptions predicting incivilities perceptions). It is estimated in Chapter 3 using panel data from Baltimore. Results show lagged, multilevel impacts of risk perceptions on changes in incivilities perceptions. Further, results show the impact of risk perceptions on seeing later changes in neighborhood problems varies significantly across street blocks. Findings support Harcourt's (2001) assertion that "disorder" is not a fixed and unambiguous label; rather, it is dependent upon a person defining his or her surroundings. People who feel a high degree of crime risk are "biased" (Hipp, 2010; Wallace, 2011) toward defining neighborhood features as more problematic than those who do not.

DEDICATION

I hereby dedicate the following work to my family. Your belief in me during the entire Graduate School process gave me the strength to complete this stage of my life. Whether it was during coursework, taking comprehensive exams, or writing the dissertation I sincerely would not have been able to do it without you.

Mom and Dad, I can't stress to you enough how many times I've reflected on how lucky I am to have you as my parents. Without your love I wouldn't have finished high school, let alone a Ph.D. You are my role models and I am very proud to say I'm becoming more like you everyday (especially Dad). I know you've both felt the stresses of my time in Grad School (especially Mom). Thank you for always going to battle for me. I love you.

Joseph, I'm the first. I know I had a head start on you, but I hope you follow me down the path, not of continuing education, but of challenging yourself. I may be receiving a Ph.D., but you are the brains of the family. Thank you for pushing me when I needed a boost and for comforting me I needed a hug. I love you. Go do great things.

Betta, thank you for all the support when writing this dissertation. You were the one that dealt with the everyday turmoil of finishing this beast. I can't thank you enough for all the time and energy you've dedicated towards helping me. You continually push me to be a better person. Thank you. I can't wait to see what the future has in store for us. I love you.

Grandma, you're my gluey rock. You hold it all together while I lean on you. Thank you for never doubting me. You inspire me to work hard and to always be a good person. You're one of the strongest women without even realizing it. The difficulties you’ve experienced and you’re grace throughtout them has taught me that I can handle any problem that comes my way. You’re an unbelievable human being. I’m so lucky to have you. I love you.

Pa, this is for you. I love and miss you everyday.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1. ACHIEVEMENT GAINS 1
    - Introduction 2
    - Review of Literature 5
    - Data 8
    - Testing the Effects on Achievement 16
        - Extension of Value-Added Models 16
        - Baseline Results 20
    - Nonrandom Classroom Assignment 27
        - Testing for High-Quality Teacher Repetition 27
        - Sorting Results 29
    - Discussion 31
    - Conclusion 39

2. NON-ACHEIVEMENT GAINS 41
    - Introduction 41

Literature Review 42

Framework for Testing Effects on Non-Achievement 44

Data 48

Results 50

Changes to Subjective Measures of Achievement 52

Changes to Students' Use of Time and Effort 56

Discussion 60

Conclusion 63

3. REVERSING BROKEN WINDOWS THEORY 65

Introduction 65

Biased Perceptions 67

Incivilities Thesis Research 68

What is "Disorder" and "Incivility"? 70

Current Investigation 72

Methods 74

Setting 74

Sample 75

Dependent Variable 77

Independent Variables 79

Covariates 80

Demographics 80

Analytic Strategy 81

Results 83

Discussion 87

Conclusions 91

BIBLIOGRAPHY 97

LIST OF TABLES

Table Page

1. Sample Summary Statistics……………………………………….……………..12
2. Summary Statistics when D_2=1………….…………………………..…..…....14
3. Summary Statistics when D=1………….……………………………..…..…....15
4. Math Scores-Fourth and Fifth Grade Samples..……………………..…..…....23
5. Reading Scores-Fourth and Fifth Grade Samples..……………………..……..24
6. Math Scores-Third, Fourth and Fifth Grade Samples..……….………..…..…...25
7. ReadingScores-Third, Fourth and Fifth Grade Samples………………..…..…...26
8. LPM Estimates-Fourth and Fifth Grade Samples..…………………..…..…...32
9. LPM Estimates-Third, Fourth and Fifth Grade Samples..……………..…..…...33
10. Non-Standardized Values of Character Trait Outcome Variables ……..…..…...50
11. Teacher Subjective Scores in Math and Reading.……………………..…..….54
12. Student Anticipated Scores in Math and Reading.……………………..…..…....55
13. Student Attendance………………………………………….……..……....57
14. Time Spent on Homework and Reading……………………………………...58
15. Summary Statistices-Individual and Street Block Variables…………..……..78
16. ANCOVA, RCR, and IASAO Models of Incivilities Perceptions……………...85
17. HLM Model of Traditional BWT model……………………………………..86
18. Fixed Effects Model of Incivilities Perceptions………………………………...88

LIST OF FIGURES

Figure Page

1. Conceptual Model ……….…………………………………..………………….73

CHAPTER 1

ACHIEVEMENT GAINS

Introduction

How long should a student and teacher maintain their educational relationship to maximize the student's human capital growth? In most U.S. public elementary schools, a student is in a teacher's classroom for 180 days (*Schools and Staffing Survey*, 2017), before moving on to the next grade and teacher. Length of school year policy and the subsequent switching of teachers is, however, arguably due more to social norms than to an analysis of what is optimal for the student. A student is best served switching to a new teacher only when the benefits outweigh the costs of doing so. For a student, benefits of the switch are associated with experiencing a new, different teacher's skill set. Opportunity costs of the switch include foregone benefits of maintaining an established relationship between the student and teacher. An increasing number of schools are questioning whether students can benefit from remaining with teachers for longer than 180 days. In both public and private schools across the U.S., looping and multi-year teaching policies are becoming more common. Programs like these allow students and teachers to remain together for multiple school years and necessitates the question of whether a student acquires more knowledge by staying with a teacher for multiple years rather than switching to a new one. This research investigates this question by comparing achievement growth in math and reading scores of elementary school students that are with a teacher for two (or more) years to "traditional" students that are only with a teacher for one year.

Looping[1] is defined as the educational practice of teachers and students remaining together for two or more years (Thompson, Franz, & Miller, 2009). Advocates of looping can be found in both peer-reviewed journal articles (Simel, 1998;Hitz, Somers, & Jenlink, 2007) and mass media outlets (Finder, 2005) alike and has led to growing implementation of this teaching practice. While many looping classrooms are not formally reported, it has been suggested that looping has been implemented, to varying degrees, in thousands of schools across the United States. In some school districts looping is offered at every grade level, while in others an individual teacher initiates the practice. Looping occurs most frequently in elementary schools, but is also present in middle schools. Both public and private schools, such as Waldorf Schools, employ looping (Grant et al., 1996). Despite the growing popularity of looping, estimates of significant effects on student achievement have not been thoroughly investigated. While this research does not specifically test the effects of looping, it does investigate whether the additional instruction time associated with looping leads to increases in a student's achievement growth. Whether a student will experience higher achievement from repeating with a previous teacher remains theoretically ambiguous.

The theoretical ambiguity of the effect on achievement largely results from the size of the adjustment costs related to a student (including parents) and teacher establishing a working, trusting relationship. Adjustment costs are mostly comprised of the time the student and teacher must spend establishing trust and becoming acquainted with each other's respective learning and teaching habits. These are paid in foregone time spent on the education curriculum and reflected by lower student achievement. Using test scores as a representative measure of human capital gain, if adjustment costs are greater than the test- score benefits of any unique skills of the new teacher, the student will score higher on math and reading achievement tests by remaining with her current teacher. When the reverse is true the student will experience higher achievement scores by moving to a new teacher's classroom.

Determining the magnitude of the adjustment costs begins with empirically testing whether students that remain with their teacher for greater than one school year experience an increased gain in achievement. Surprisingly, there is little peer-reviewed research on this seemingly interesting topic. The limited amount that is available comes from Cistone and

Shneyderman (2004) and Rodriguez and Arenz (2007). In both papers, the authors find that looped elementary students perform significantly better than their non-looped counterparts in measures such as math and reading achievement, writing strategies, and vocabulary.

While these descriptive results provide preliminary evidence regarding the benefits of students staying with teachers for multiple years, by simply comparing mean outcomes of groups of students that have been looped with those that have not, they lack controls for confounding influences that can create bias in the estimated effects. Such confounding influences include student-specific characteristics, teacher-specific characteristics including teacher quality, and self-selection into the treatment group. By controlling for these con- founding variables, the current research enlightens this darkened corner of the literature.

To answer this question, I extend the Value-Added Models (VAM) of teaching quality initially proposed by E. Hanushek (1971) and later refined by Todd and Wolpin (2003), Rockoff (2004), and Rivkin, Hanushek, and Kain (2005) among many others. While widely used, these models of teacher quality are not without critics. Jesse Rothstein highlighted the bias in VAM estimates associated with nonrandom student sorting into teachers' class- rooms (Rothstein, 2009, 2010). However, since Rothstein's criticisms, researchers have found new ways to mitigate the bias in VAM estimates to insignificance by evaluating teachers over multiple years (Koedel & Betts, 2011) and by including appropriate controls in the VAM framework (Chetty, Friedman, & Rockoff, 2014b). Both of these techniques are used in the current investigation.

Much of the ongoing dispute over the validity of VAMs is centered around the magnitude of bias in their estimates. While this is a necessary and important concern, it overshadows the usefulness of VAMs. VAMs can be used to control for teacher quality while investigating the impacts of other educational inputs. In this research, I use the underlying framework of VAMs to control for teacher quality. This allows me to investigate the causal effects of repeating with a previous teacher independent of a teacher's quality.

In doing so I underscore, despite the flaws commonly associated with VAMs, they are still one of the most helpful tools education researchers can use to control for teacher quality.

Controlling for teacher quality proves to be very important in the current research. Using matched student and teacher data from the North Carolina Education Research Data Center (NCERDC), this research begins with an Ordinary Least Squares regression model to test whether a student experiences achievement gains from repeating with a teacher he has had previously. Initially, while only controlling for student, school, and classroom/peer characteristics I find results that agree with Cistone and Shneyderman (2004) and Rodriguez and Arenz (2007). Specifically, I find that students experience statistically significant increases in achievement scores in both reading and math. However, once I control for teacher quality using a VAM structure where past inputs decay geometrically, the positive estimates of repeating with a teacher on math and reading test scores are almost completely mitigated and no longer statistically different from zero.

This latter finding highlights that *not* controlling for teacher quality may create upward omitted variable bias in estimates measuring the effects of repeating with a teacher on student achievement. The initial upwardly-biased estimates are due to a positive correlation between the variable of interest (whether a student is repeating with a previous teacher or not) and the omitted variable (teacher quality). I hypothesize that this is most likely associated with higher quality teachers having a larger proportion of repeating students than lower quality teachers. I investigate this assertion, by using a number of limited dependent variable models to test whether higher quality teachers have an increased probability of having a student she has previously taught. I find support for this hypothesis by estimating that higher quality teachers have a 29% to 46% greater chance of having a repeating student.

The research conducted in this investigation emphasizes that controlling for a teacher's quality is paramount to understanding the payoffs of a student repeating with a teacher. Accounting for teacher quality, something not done in previous studies, yields surprising and opposing results. While this research does not investigate all the effects of repeating with a teacher and students may still benefit in non-academic measures, it does gives strong evidence that repeating with a teacher doesn't academically benefit the student as much as initially determined.

The rest of this paper is organized as follows: The second section reviews the literature, the third specifies the data used to answer the research questions, the fourth tests the effects of repeating with a teacher on student achievement, the fifth tests nonrandom classroom sorting, the sixth discusses the results and the seventh concludes and offers future work recommendations.

Review of Literature

Attempts to identify the true effects teachers have on student achievement are well-documented (Gordon, Kane, & Staiger, 2006; Lockwood & McCaffrey, 2009) Almost all of the research in this area, however, focuses on a teacher's yearly impact on a student. There is little published literature on the effect of a student having a teacher for more than a year and specifically whether the relationship between the time a student is in a teacher's classroom exhibits constant, increasing or decreasing returns to academic achievement.

Before theorizing about the relationship between student achievement and time in a teacher's classroom, it is necessary to understand the overarching effects teachers have on student academic achievement and the importance of allocating teachers to students efficiently. Rivkin et al. (2005) find, while controlling for school, grade, and student het- erogeneity, that a one standard deviation increase from the mean in teacher quality raises average student achievement by about .1 standard deviations of the standardized math test score distribution. This result highlights that a 10-student reduction in class size has a smaller positive effect on achievement than a one standard deviation increase in teacher quality. E. A. Hanushek (2003) highlights the importance of teacher quality as well as the quality of education overall. He reviews the impacts of the dramatic increases in school resources, such as smaller class sizes and tighter teacher credentials, during the 1990s and early 2000s. He compares these estimated effects with an incentive system designed to increase teacher quality, finding that the change to the incentive system is almost always superior to an increase in educational inputs.

The number of years a student spends in primary education has long been a common proxy for human capital formation. While the quantity of educational attainment provides basic but important information on an individual's productivity, the quality of the education attained should not be overlooked. (Dugan, Jamison, & Radner, 1976; Manski & Wise, 1983; Rivkin, 1995) find that a student's education quality has significant effects on both the quantity of education attained (College/high school completion) and on the individual's productivity capabilities. Furthermore, Chetty, Friedman, and Rockoff (2014c) show that estimates of teacher impacts on test scores (as measured by Value-Added Models) are a good measure of teacher quality and 2) students assigned to high quality teachers are more likely to attend college, earn high salaries, and are less likely to have children as teenagers. Given the benefits students experience from having a high-quality teacher, it is critical to understand how to best allocate teachers in order to maximize the level of teaching quality students receive. Previous literature tests for changes in teacher quality when students and teachers are matched based on gender, race, and ethnicity, but finds no significant statistical differences (Ehrenberg, Goldhaber, & Brewer, 1995; Winters, Haight, Swaim, & Pickering, 2013). Kenny (1982) provides a theoretical framework to measure education input quality and finds that "schools operate in the range of increasing returns to school inputs." Kinsler (2016) extends the basic Value-Added Model research to include teacher complementarities across grades. Although Kinsler finds only very small instructional complementarities between different teachers, he does not explore whether there are instructional complementarities across grades for the same teacher with the same student. The current research investigates opportunities to increase teacher quality by reallocating teachers to classrooms containing students they've previously taught.

To best explain how teacher reallocation can present opportunities for increased teacher quality, it is useful to understand the evolving relationship between teachers and students. From the School Psychology literature, Simonds et al. (1994) find that both students and teachers recognize that interpersonal communication is key to the academic success of the student-teacher relationship. Shamai, Ilatov, Hertz-Lazarovitz, and Bentsvi-Mayer (1995) establish that the teacher bears more responsibility for establishing and maintaining this relationship. This responsibility, or "power to realize" (Campbell, Kyriakides, Muijs, & Robinson, 2003) emphasizes that effective teachers have "power" over what occurs in their classroom. Kerssen-Griep, Gayle, and Preiss (2006) extend this result by concluding that this power is most likely derived from the authority teachers have in their classrooms and that students give this authority to teachers by consenting and cooperating towards the common goal of education attainment (Brown, 2004; Elliott, 2009). The time a teacher needs to establish classroom authority and the length of time this authority continues is difficult to quantify. Empirically testing the achievement differences between classrooms where the teacher's authority is already realized (classroom where an existing student-teacher relationship is established) and classrooms where it is not, is an important first step in understanding the opportunity cost of establishing classroom authority.

Students that begin the school-year in the classroom of a teacher they've had previously and where teacher authority has already been established, provides teachers with additional time to allocate towards instruction. Teachers estimate they gain an additional month of instructional time during the second year repeating with students (Gaustad, 1998). Furthermore, this affords additional stability, enhancing relationships between the teacher and the student and the teacher and the student's parents. The additional stability also pro- vides the opportunity for deeper student engagement in the education process (Thompson et al., 2009). Enhanced student engagement has been found to be especially true for students whose lives outside of school are unstable (Simel, 1998).

Repeating with a teacher is not without its disadvantages, however. In classrooms where classmates are with a teacher for a second year, there is a greater chance for a student that has not had a teacher previously to become jealous. Because teachers are required to prepare lessons that cover at least two years' worth of material, they have less time to perfect their lessons through repetition. Additionally, if a teacher has any specific weaknesses in her teaching skill set, these weaknesses will persist for multiple years. Similarly, teachers have concerns about being placed with a difficult student and/or parents for multiple years. (Thompson et al., 2009).

Empirically understanding whether the benefits of repeating with a teacher outweigh its costs is difficult because students are usually non-randomly selected to repeat with a teacher. Students that have exceptional learning needs are specifically targeted as well as teachers that volunteer for looping teaching assignments. Because these students and teachers are not random draws from the overall school population, accurately comparing student achievement is difficult. To alleviate this problem and to drastically increase the sizes of the samples used in this research, I use information on students matched to past and present teachers using a large dataset from the North Carolina Education Research Data Center.

Data

Information on a large number of students' test scores and educational inputs must be accurately matched to their classroom assignments and teachers' characteristics over multiple years to determine whether students are repeating with a previous teacher. Students have anywhere from a 2%-4% chance of being taught by a teacher they've had previously. In order for estimates on any student achievement effects from repeating with a teacher to be precise enough and given the low rate of student-teacher repetition, a large original sample of students is needed. Fortunately, the North Carolina Education Research Data Center (NCERDC) provides this type of data. The NCERDC has developed a longitudinal dataset that tracks achievement and characteristics of students as they progress through each year in the entire state school system of North Carolina. The dataset also includes yearly information on the specific teacher, class, and school to which the students are assigned.

I obtained information on the school years ranging from 1993-1994 to 2012-2013(for simplicity, henceforth, each school year is identified by the year in which the spring semester occurs, e.g. the school year 2012-2013, is 2013). Information on teacher class- room assignment is not specifically identified in years before 2006[2]. Therefore, I restrict the dataset for this investigation to cover the years from 2007 to 2013. Because this investigation is focused on elementary school students, I limit the initial dataset used to find whether a student has repeated with a teacher to grades 1 through 5. However, the initial dataset is reduced to include only grades 3 through 5 because test scores are unavailable for students in grades below third grade,

The dataset from the NCERDC provides an abundance of information on a number of educational input variables. This includes a student's age, ethnicity, gender, exceptionality code[3], free/reduced lunch, days absent, limited English proficiency, and academic giftedness. Information on measures of educational inputs at the teacher, classroom, and school levels are also included in the dataset.

I restrict the sample to include students assigned to self-contained classrooms[4]. Observations that were updated, but where the original observation was not deleted, were collapsed into one observation. This led to eliminating about 42% of the initial (grades 1 through 5) dataset[5], resulting in 3,435,774 remaining observations. Of these remaining observations, only third to fifth grade students with non-missing test scores in math or reading are retained. The resulting testable population contains 1,705,626 observations. I stop at grade 5 because the number of students assigned to self-contained classes is significantly smaller in later grades. This is because students are more likely to switch classes and there- fore have different teachers for different subjects. In order to not attribute one teacher's effect for another's, 6th grade students and higher are excluded.

The samples used for estimation are restricted in a number of ways. First, because estimation of teacher effects is conditioned on prior test scores, the samples are limited to students that have prior-year's test scores available. Second, students with missing information on their classroom teacher and on the school they attended are excluded. Third, I limit the samples to classrooms with sizes more than 7 students and less than 45. Fourth, I eliminate students from charter schools because this research is interested in gains in achievement in public schools. Finally, observations that are missing information on individual characteristics, such as whether the student receives a free/reduced lunch or whether the student is academically gifted, are eliminated from the samples.

These sample restrictions lead to a loss of a large number of observations from the testable population. The majority of this reduction in observations is due to the prior-year's test score restriction. This restriction, as well as the non-missing information requirement, results in higher-achieving students remaining in the samples used for estimation because lower-achieving students are more likely to have missing information across years. However, comparing the population to the samples used for estimation in Table 1 there are limited differences between student summary statistics of the population and of the samples. With the exception of an increased percentage of academically gifted students in the samples, sample averages of student covariates are similar to the population's. This provides evidence that the samples are random and representative draws of the population. Regression analysis of math and reading test scores is done using 2 overlapping samples. The first, includes only fourth and fifth grades while the second consists of third, fourth, and fifth grades. I do this for two reasons; 1) As mentioned above, test scores prior to grade 3 are unavailable. However, math and reading tests administered at the beginning of the third grade (third grade pre-tests) are available.

These tests can serve as a proxy for second grade achievement and therefore can be used as a control for assessing achievement gains in third grade. This increases the number of observations available for analysis and follows previous research (Rothstein, 2009). Unlike the second grade proxies,, third and fourth grade test scores are administered at the end of the year and are used as lagged controls for fourth and fifth grade achievement, respectively. I use separate samples to identify any differences in estimates associated with the inclusion of second grade test score proxies. Third grade pre-test scores include any learning depreciation that occurs over the summer between second and third grades that would otherwise be absent in second grade end-of-year test scores. 2) Information on second grade students' absences is unavailable. Previous literature that tests for teacher effects on student achievement typically controls for prior-year's absences (Chetty et al., 2014b). While I use this approach in the fourth and fifth grade samples, because of the unavailability of second grade absences, I use current year absences as a control in the third through fifth grade samples. In the fourth and fifth grade samples, I eliminate students who were absent for more than one-third of the *previous* year. In the third through fifth grade sample, I eliminate students who were absent for more than one-third of the *current* year. This sample restriction helps to eliminate any outlying observations that have been abnormally affected by extraneous events.

The summary statistics in Table 1 show that expanding the repeated cross-sectional samples to include third grade students, increases the number of observations by 160,683 for math scores and 167,381 for reading. The number of students increases by 43,968 for math and 45,397 for reading. In the third, fourth, and fifth grade samples, current and lagged test scores and lagged classroom-averaged test scores are slightly lower than in the fourth and fifth grade sample. The majority of individual student variables are similar both between the two samples and in comparison to the testable population. There are a few exceptions that warrant explanation. The proportion of third grade students in the third, fourth, and fifth grade samples is significantly lower than in the population. Because of

Table 1
Sample Summary Statistics

Variables	Population	4th/5th Sample Math	4th/5th Sample Reading	3rd/4th/5th Sample Math	3rd/4th/5th Sample Reading
Observations	1,705,626	538,094	534,253	698,777	701,634
Students	878,646	376,028	373,486	419,996	418,883
Schools	1,431	1,307	1,306	1,356	1,355
Teachers	29,686	14,606	14,594	20,890	21,006
Student Statistics					
Std. Math Score	0.000767	0.0759	–	0.0636	–
	(1.00)	(1.00)	–	(0.997)	–
Std. Reading Score	-0.00347	–	0.0604	–	0.0508
	(1.00)	–	(0.996)	–	(0.995)
Lagged Std. Math Score	0.0199	0.0999	–	0.0713	–
	(0.993)	(0.991)	–	(0.999)	–
Lagged Std. Reading Score	0.0112	–	0.0891	–	0.0649
	(0.994)	–	(0.988)	–	(0.996)
Female	49.41%	49.59%	49.76%	49.61%	49.77%
Age (years)	10.31	10.83	10.82	10.51	10.50
Grade 3	35.69%	–	–	21.81%	22.72%
Grade 4	33.49%	52.07%	52.08%	40.64%	40.17%
Grade 5	30.82%	47.93%	47.92%	37.54%	37.10%
Black	26.00%	24.82%	24.83%	24.84%	25.04
Hispanic	12.87%	12.60%	12.52%	12.31%	12.26%
Other nonwhite	7.99%	7.86%	7.85%	7.86%	7.86%
Free/Reduced Lunch	52.24%	50.40%	50.34%	50.16%	49.96%
Academically Gifted	14.49%	20.57%	20.70%	17.39%	17.60%
Exceptional Student	12.78%	12.63%	12.07%	12.80%	12.30%
Limited English Proficiency	10.11%	10.11%	9.99%	9.92%	9.75%
Days in School	161.51	161.47	161.48	164.09	164.12
Lagged Days in School	160.52	163.66	163.66	160.59	160.65
Moved During School Year	0.903%	0.590%	0.588%	0.089%	0.087%
Repeating Grade	1.30%	0.891%	0.888%	1.67%	1.64%
With Teacher 2nd Time($D_2 = 1$)	40,927	18,646	18,499	21,760	21,553
Repeating with Teacher($D = 1$)	44,853	19,161	19,002	22,452	22,234
Classroom Statistics					
Number Students in Class	21.75	23.04	23.05	22.43	22.42
L. Class Ave.-Math	0.00694	0.0398	–	0.0317	–
L. Class Ave.-Read	0.000723	–	0.0321	–	0.0233
Ave. Teacher V.A. Estimate	–	0.0004	0.0002	0.0009	0.0004
		(0.236)	(0.170)	(0.239)	(0.175)

Note: Data was made available by the NCERDC. Population is based on students in third, fourth, and fifth grades in self-contained classrooms. Standard errors are in parentheses. Lagged test scores include beginning-of the-year pre-tests that serve as proxies for second grade test scores.

missing data constraints, only $\approx$ 44% of the population's original 608,738 third grade observations have third grade pre-test information. There are no apparent trends or patterns to which observations include data on third grade pre-test scores and given the similarities in proportions of the other individual covariates, it is assumed that the students with non-missing third grade pre-test information have been randomly selected. Another difference between the samples and population is the number of academically gifted students is proportionately greater in both samples than in the population. This is due to the in- creased likelihood of higher achieving students remaining in the sample. This is an issue I address, by additionally testing regressions that exclude gifted students[6] finding similar results across all samples.

To determine whether a student has had a teacher before, a simple dummy variable, D_n, was constructed identifying whether a student was previously taught by their current teacher. Current-year, student-teacher matches were cross-referenced with a student's previous teachers for up to the past 4 years. Specifically, the teacher code matched with a student in grade 5 was cross-referenced with the teacher codes matched with the same student in grades 1 through 4. I match fourth grade students' contemporaneous teacher codes with their previous teacher codes from grades 1 through 3 and the teacher codes for students in grade 3 were matched with their previous teacher codes from grades 1 and 2. For students that are currently being taught by a teacher they have had previously, $D_n = 1$. For all other students, $D_n = 0$, indicating that these students are being taught by a teacher for the first time. 2.40% of the population repeat with a teacher for a second time ($D_2 = 1$) and 2.63% of the population are in a teacher's classroom that has taught them before ($D = 1$)[7]. In both the math and reading samples, these percentages are slightly higher. In the fourth and fifth grade samples, students repeat with a teacher for a second year 3.46% of the time in both math and reading. In the third through fifth grade samples,

Table 2

Summary Statistics when $D_2 = 1$

	4th/5th Sample		3rd/4th/5th Sample	
Variables	Math	Reading	Math	Reading
Test Score Observations	18,646	18,499	21,760	21,553
Students	18,640	18,493	21,635	21,428
Schools	1,101	1,097	1,179	1,173
Teachers	3,136	3,121	4,119	4,087
Student Statistics				
Std. Test Score	0.1407	0.1119	0.1043	0.07155
	(1.01)	(0.996)	(1.00)	(0.998)
Lagged Std. Test Score	0.1472	0.1247	0.0718	0.0508
	(1.01)	(1.00)	(1.04)	(1.03)
Female	50.02%	50.20%	49.87%	50.03%
Age	10.92	10.92	10.76	10.76
Grade 3	–	–	12.66%	12.55%
Grade 4	43.60%	43.58%	38.08%	38.12%
Grade 5	56.40%	56.42%	49.26%	49.33%
Black	24.63%	24.66%	25.36%	25.42%
Hispanic	13.35%	13.27%	13.67%	13.58%
Other nonwhite	7.81%	7.81%	7.63%	7.62%
Free/Reduced Lunch	49.91%	49.71%	51.41%	51.26%
Academically Gifted	23.41%	23.56%	20.92%	21.10%
Exceptional Student	13.94%	13.32%	14.98%	14.38%
Limited English Proficiency	10.80%	10.65%	11.45%	11.28%
Days in School	163.49	163.51	163.96	163.96
Lagged Days in School	163.60	163.60	161.05	161.11
Moved During School Year	0.193%	0.184%	0.060%	0.065%
Repeating Grade	2.91%	2.89%	6.93%	6.90%
2nd Year with Teacher	18,646	18,499	21,760	21,553
Classroom Statistics				
Number Students in Class	22.44	22.45	22.11	22.13
Lagged Class Average	0.0424	0.0304	0.0304	0.0176
	(0.415)	(0.397)	(0.413)	(0.396)
Ave. Teacher V.A. Estimate	0.00677	0.00934	0.00574	0.00856
	(0.229)	(0.180)	(0.237)	(0.196)

Note: Data was made available by the NCERDC. Population is based on students in third, fourth, and fifth grades in self-contained classrooms where current-year test score information is available. Standard errors are in parentheses. Lagged test scores include beginning-of the-year pre-tests that serve as proxies for second grade test scores.

Table 3

Summary Statistics when $D = 1$

	4th/5th Sample		3rd/4th/5th Sample	
Variables	Math	Reading	Math	Reading
Test Score Observations	19,161	19,002	22,452	22,408
Students	18,941	18,782	21,978	21,764
Schools	1,103	1,099	1,181	1,175
Teachers	3,164	3,148	4,156	4,123
Student Statistics				
Std. Test Score	0.1366	0.1086	0.1014	0.0708
	(1.01)	(0.998)	(1.00)	(1.00)
Lagged Std. Test Score	0.1428	0.1223	0.0701	0.0518
	(1.01)	(1.00)	(1.04)	(1.03)
Female	49.88%	50.07%	49.84%	50.00%
Age	10.92	10.92	10.75	10.75
Grade 3	–	–	13.07%	12.97%
Grade 4	43.63%	43.60	37.92%	37.95%
Grade 5	56.37%	56.40	49.01%	49.08%
Black	24.83%	24.85%	25.61%	25.67%
Hispanic	13.40%	13.33%	13.73%	13.65%
Other nonwhite	7.83%	7.83%	7.67%	7.67%
Free/Reduced Lunch	50.17%	49.96%	51.50%	51.34%
Academically Gifted	23.29%	23.46%	20.88	21.05%
Exceptional Student	14.22%	13.57%	15.21%	14.59%
Limited English Proficiency	10.87%	10.72%	11.51%	11.35%
Days in School	163.46	163.48	163.93	163.93
Lagged Days in School	163.59	163.60	161.09	161.15
Moved During School Year	0.266%	0.247%	0.125%	0.121%
Repeating Grade	2.91%	2.89%	6.83%	6.79%
Repeating with Teacher	19,161	19,002	22,452	22,234
Classroom Statistics				
Number Students in Class	22.40	22.41	22.07	22.09
Lagged Class Average	0.0395	0.0289	0.0289	0.0182
	(0.415)	(0.397)	(0.414)	(0.399)
Ave. Teacher V.A. Estimate	0.00575	0.00846	0.00487	0.00774
	(0.229)	(0.181)	(0.237)	(0.197)

Note: Data was made available by the NCERDC. Population is based on students in third, fourth, and fifth grades in self-contained classrooms where current-year test score information is available. Standard errors are in parentheses. Lagged test scores include beginning-of the-year pre-tests that serve as proxies for second grade test scores.

students repeat with a teacher for a second year 3.07% (Reading) and 3.11% (Math) of the time. These figures in- crease by 0.10% in all samples when calculating the ratio of observations that repeat with a teacher any number of times. The summary statistics when $D_2 = 1$ and $D = 1$ are provided in Tables 2 and 3, respectively.

Testing the Effects on Achievement

Extension of Value-Added Models

Using the initial framework proposed by E. Hanushek (1971) and Boardman and Murnane (1979), this research builds on the Value-Added Models (VAMs) highlighted in Sass, Semykina, and Harris (2014) and Chetty et al. (2014b). VAMs serve as an empirically measurable production function of a student's knowledge. These models assume a student's accumulated knowledge is determined by a production process in which current and past time-varying inputs are combined with a student's time-invariant inputs. A student's time-invariant inputs are primarily comprised of a student's innate ability. It also includes individual and family inputs that do not vary over time. The VAMs used in this research generalize the time-varying inputs into two vectors of characteristics; those that are individually based (including family inputs) and those that are based on the student's educational institution. Let $\boldsymbol{X}_{it}$ and $\boldsymbol{W}_{it}$ represent the history of individual and school-based educational inputs, respectively. The student's time-invariant component is denoted μ_i. Allowing for idiosyncratic error in achievement, ε_{it}, the production function of achievement at time, t, is

$$A_{it} = A_t[\boldsymbol{X}_{it}, \boldsymbol{W}_{it}, \mu_i, \varepsilon_{it}] \tag{1}$$

Where the subscript on $A_t[.]$ allows the impacts of inputs and ability to vary throughout time.

$\boldsymbol{W}_{it}$, the vector of educational inputs, includes both school-level and classroom-level inputs. School-level inputs include such variables as principal quality and the effectiveness of the school's administrative staff. Classroom-level inputs include peer-effects of the students in the classroom[8], the quality of the teacher, and the primary variable of interest, a variable indicating whether or not a student and teacher have been paired together previously.

Assuming the cumulative achievement process is linear and additively separable, and explicitly partitioning out D_{it}, the input variable indicating that a student has been matched with her current teacher previously, from $\boldsymbol{W}_{it}$, equation (1) can be written

$$A_{it} = \sum_{h=0}^{t} [\boldsymbol{\alpha}_{ht}\boldsymbol{X}_{ih} + \boldsymbol{\beta}_{ht}\boldsymbol{E}_{ih} + \delta_{ht}D_{ih}] + \psi_t\mu_i + \varepsilon_{it} \tag{2}$$

The above equation specifies that achievement at time t is determined by the accumulation of all past and present, time-varying, student- and school-based inputs; time-invariant student inputs; and idiosyncratic student-specific shocks. The (potentially) time-varying vectors $\boldsymbol{\alpha}_{ht}$ and $\boldsymbol{\beta}_{ht}$ represent the effects on achievement associated with individual and school inputs, respectively. δ, the average treatment effect of having a teacher for a previous year, is the parameter of interest. The impact of μ_i, the fixed student component for individual i, is given by ψ_t and can vary over time. Furthermore, the achievement function is time-specific and allows the marginal effects of the independent variables to vary at each grade.[9]. Estimation of equation (2)'s parameters measures the average treatment effect for each educational input within a specific year-grade combination[10].

Despite its flexibility and lack of restrictive assumptions, there is a high computational cost of measuring the coefficients in equation (2). To do so, a researcher needs information on all individual and school inputs from every year (including the current year) the student has been acquiring knowledge. Given this unrealistically onerous task, the current research follows previous VAM literature by assuming the achievement function does not vary with grade level. Specifically, the impact of an input on achievement varies only with the time span between the application of the input and the measurement of achievement, but does not vary with the grade level at which it is applied. Under this assumption for example, the effect on a student's third grade achievement from having a small class in first grade is the same as the effect on sixth grade achievement from having a small class in fourth grade. Formally, for any t:

$$A_{it} = \sum_{h=0}^{t} [\boldsymbol{\alpha}_h \boldsymbol{X}_{i,t-h} + \boldsymbol{\beta}_h \boldsymbol{E}_{i,t-h} + \delta_h D_{i,t-h}] + \psi_t \mu_i + \varepsilon_{it} \tag{3}$$

Where α_h, β_h, and δ_h vary only with the time span of input application, h, while ψ_t still has the flexibility to vary over time.

It is also assumed that the marginal impacts of prior educational inputs decay geometrically between the time of an input's application and the time at which achievement is measured. Specifically, inputs decay at a rate, λ, where $0 \leq \lambda \leq 1$, such that for any h, $\boldsymbol{\alpha}_{t-h} = \lambda \boldsymbol{\alpha}_{t-1-h}$, $\boldsymbol{\beta}_{t-h} = \lambda \boldsymbol{\beta}_{t-1-h}$, and $\delta_{t-h} = \lambda \delta_{t-1-h}$. Imposing this assumption, equation (3) becomes:

$$A_{it} = \sum_{h=0}^{t} \lambda^h [\boldsymbol{\alpha} \boldsymbol{X}_{i,t-h} + \boldsymbol{\beta} \boldsymbol{E}_{i,t-h} + \delta D_{i,t-h}] + \psi_t \mu_i + \varepsilon_{it} \tag{4}$$

Adding and subtracting, a student's decayed prior year's achievement, where $\lambda A_{i,t-1} = \lambda \left(\sum_{h=0}^{t-1} \lambda^h [\boldsymbol{\alpha} \boldsymbol{X}_{i,t-1-h} + \boldsymbol{\beta} \boldsymbol{E}_{i,t-1-h} + \delta D_{i,t-1-h}] + \psi_{t-1} \mu_i + \varepsilon_{i,t-1} \right)$, from both sides of (4) yields:

$$\begin{aligned} A_{it} = & \left(\sum_{h=0}^{t} \lambda^h [\boldsymbol{\alpha} \boldsymbol{X}_{i,t-h} + \boldsymbol{\beta} \boldsymbol{E}_{i,t-h} + \delta D_{i,t-h}] + \psi_t \mu_i + \varepsilon_{it} \right) + \lambda A_{i,t-1} - \\ & \lambda \left(\sum_{h=0}^{t-1} \lambda^h [\boldsymbol{\alpha} \boldsymbol{X}_{i,t-1-h} + \boldsymbol{\beta} \boldsymbol{E}_{i,t-1-h} + \delta D_{i,t-1-h}] + \psi_{t-1} \mu_i + \varepsilon_{i,t-1} \right) \end{aligned} \tag{5}$$

Collecting and simplifying like terms yields:

$$A_{it} = \boldsymbol{\alpha X_{it}} + \boldsymbol{\beta E_{it}} + \lambda A_{i,t-1} + \delta D_{it} + (\psi_t - \lambda \psi_{t-1})\mu_i + \eta_{it} \tag{6}$$

Where $\eta_{it} = \varepsilon_{it} - \lambda \varepsilon_{i,t-1}$. Equation (6) serves as the baseline model for this research. (6) states that current achievement is a function of four components: current student-based and education-based inputs, including D_{it}; decayed lagged achievement; student-specific heterogeneity; and any autocorrelated, idiosyncratic shocks. The use of $A_{i,t-1}$ introduces autocorrelation in the error terms, but its inclusion is for good reason. $A_{i,t-1}$ serves as an approximation for all past individual and educational inputs, circumventing the need for a full history of those measures. Furthermore, the inclusion of a student's lagged achievement, at least partially, accounts for student heterogeneity.

In order to use the baseline model found in (6) as a testable hypothesis, two final assumptions about the decay rate of educational inputs, λ, and student heterogeneity, μ_i, are made. First, it is assumed that the decay rate of educational inputs is such that $0 < \lambda < 1$. This assumption follows previous literature[11]. Second, the use of $A_{i,t-1}$ and student covariate controls is assumed to completely control for student heterogeneity, such that $\psi_t = \lambda \psi_{t-1}$ dropping the term $(\psi_t - \lambda \psi_{t-1})\mu_i$ from equation (6).[12]

δ serves as a measure of the student achievement growth associated with being a student of a teacher for longer than 1 school-year. This treatment effect is independent of teacher quality, school quality, and student characteristics. It measures whether the increased instructional time associated with a teacher and student having an established relationship has an impact on a student's achievement in math or reading. In this investigation, estimates of δ are found for students repeating with a teacher for a second year only ($D_2 = 1$) and for students repeating with a teacher anywhere from a second to fourth year ($D = 1$).

Baseline Results

To comprehensively understand the value of δ from equation (6), I estimate the effect of repeating with a teacher using a number of regressions. I find estimates using both the fourth and fifth grade samples (Results found in Tables 4 and 5) and the third through fifth grade samples (Results found in Tables 6 and 7). I use math test scores and reading test scores as separate dependent variables for each regression. Results for estimates using math test scores are reported in Tables 4 and 6 and results for estimates using reading test scores are reported in Tables 5 and 7. In Tables 4 through 7, estimates of δ_2 and δ represent the effects of having a teacher for a second year ($D_2 = 1$) and of repeating with a teacher($D = 1$). In all regressions, standard errors are clustered at the school level. Following the work of Chetty et al. (2014b), in column (1) of tables 4 through 7 only cubic polynomials of prior-year achievement in math and reading are included as regressors in estimation. Under the assumptions above, they are used as controls for all prior educational inputs. Estimates of δ_2 and δ are significant across both subjects and samples.

In (2) of Tables 4 through 7, I extend the estimation of the δs by adding student co-variates to regressions (1). These include controls for a student's gender, age[13], ethnicity[14], economically disadvantaged background, academic giftedness, exceptional condition (see footnote 3 for further explanation), limited English proficiency, attendance[15], switching of schools during the year[16], and whether the student is currently repeating a grade. (2) offers the most similar replication to the descriptive results found in Cistone and Shneyderman (2004).

The δ estimates for reading test scores in column (2) are about half of the δ estimates for math scores from their respective samples. Math δ estimates are significant at the 99% level and range from 0.015 to 0.017 of a standard deviation increase in math achievement. Reading δ estimates are just barely significantly different from zero with estimates ranging from 0.007 to 0.009 of a standard deviation increase in reading achievement.

The significance of the δ estimates in reading is sensitive to the inclusion of school-level clustered standard errors and to the sample used for estimation.

Columns (1) and (2), despite not including many necessary education inputs, emphasize a strong possibility of a positive increase in math achievement associated with repeating with a teacher. A similar, but weaker possibility also exists for reading achievement. Further estimation that includes controls for school-level inputs, such as the quality of the school the student attends, the student's classmates, and the student's teacher is necessary.

In (3) and (4), time, grade, and school fixed effects are added to the previous regressions. Time and grade fixed effects are interacted with one another and a student's school is indicated via a dummy variable that captures the effects on achievement associated with the composition of the school. This includes the school's principal, any administrative procedures, and interaction of the students (school-level peer effects).

To account for classroom-level peer effects, controls for the number of students in the classroom and lags of the class averages in math and reading test scores[17] are included in (5). The resulting estimates of δ_2 and δ are found by contrasting achievement of students that repeat with a teacher, to students with comparable backgrounds, in the same school and grade, with similar classmates, during the same year that do not repeat with a teacher.

In math, the magnitude of the estimates of δ_2 and δ varies with the subsequent addition of independent variables in regressions (1) through (5), however, in both samples, estimates of the δs are found to be significantly different from zero. δ estimates from the fourth and fifth grade sample are slightly more positive than those from the third through fifth grade sample. In (5), while measuring all feasible educational inputs excluding those directly associated with the teacher, the estimated increase in math achievement of repeating with a teacher is 0.014-0.016 standard deviations of the test score distribution.

To put the δ measures into context, previous estimates of the increase in a student's math achievement from having a "high" quality teacher [18] range from 0.11 to 0.26 standard deviations of the test score distribution[19]. While restricting the estimation of teacher effects to teachers with more than 20 student observations (across all years of the sample), I find estimates on math achievement to be 0.16 standard deviations in the grades 4 and 5 sample and 0.18 standard deviations in the grades 3 through 5 sample. Using these values, repeating with a teacher seemingly has anywhere from 7.5% to 10% of the effect of having a high quality teacher.

In reading, estimates of the δs from regressions (3) through (5) are also positive, however, they are mostly not significantly different from zero in both samples. Similar to the math regressions, the magnitude of the δ reading estimates in the fourth and fifth grade sample ($\approx$0.007 reading test score standard deviations) is greater than the magnitude of the same measures in the third through fifth grade sample ($\approx$0.005 S.D.). The lack of significance in these estimates is not entirely surprising. While past research has shown that high quality teachers have a positive effect on reading achievement (Rockoff, 2004; Rivkin et al., 2005; Kane, Rockoff, & Staiger, 2008; B. A. Jacob & Lefgren, 2008; Rothstein, 2010), that effect is estimated to be smaller than the estimated effect on math achievement of the same teacher quality increase. In both reading samples of the current research, the increase in reading achievement from having a high quality teacher is estimated to be about 0.12 standard deviations of the reading rest score distribution[20]. This finding falls within the range of similar estimates (0.08 to 0.18 S.D.) from past research (E. A. Hanushek & Rivkin, 2010). The estimates in column (5) from Tables 5 and 7 make up 4% to 6% of the high-quality teacher effects on reading achievement. Despite the lack of statistical significance, these estimates on reading achievement serve as a starting point in understanding whether estimates of the effect of repeating with a teacher are sensitive to the inclusion of teacher quality controls.

In regression (6), using an indicator variable for each teacher id, teacher fixed effects are included to control for teacher quality during estimation. Interestingly, the esti-

Table 4

Math Scores-Fourth and Fifth Grade Sample

Regression	(1)	(2)	(3)	(4)	(5)	(6)
Estimate of δ_2	0.02179***	0.01729***	0.01585***	0.01594***	0.01606***	0.004029
S.E. of δ_2	(0.00346)	(0.00389)	(0.00388)	(0.00392)	(0.00391)	(0.00433)
Estimate of δ	0.02060***	0.01596***	0.01224***	0.01595***	0.01572***	0.004816
S.E. of δ	(0.00341)	(0.00384)	(0.00384)	(0.00388)	(0.00387)	(0.00431)
Prior-Year Test Score	Y	Y	Y	Y	Y	Y
Student Demographics		Y	Y	Y	Y	Y
Grade and Year F.E.s			Y	Y	Y	Y
School F.E.s				Y	Y	Y
Classroom Characteristics					Y	Y
Teacher F.E.s						Y
number of observations	755,764	538,321	538,321	538,321	538,321	538,321
R-squared	0.7080	0.7303	0.7310	0.7417	0.7424	0.7710

Note: $\delta 2$ is an estimate of the effect of repeating with a teacher on achievement when the student repeats with a teacher for a second time only. Where as, δ is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third, or fourth time. Lagged test scores consist of cubic-polynomials of math and reading scores. Classroom characteristics include number of students and cubics of lagged average classroom achievement in math and reading. *, **, *** represents significance at the 90, 95, and 99% levels, respectively

Table 5
Reading Scores-Fourth and Fifth Grade Sample

Regression	(1)	(2)	(3)	(4)	(5)	(6)
Estimate of δ_2	0.009049***	0.008583**	0.007065*	0.006653*	0.006649	-0.004525
S.E. of δ_2	(0.00352)	(0.00398)	(0.00398)	(.00406)	(.00406)	(0.00464)
Estimate of δ	0.008325**	0.008791***	0.007286*	0.006601*	0.006503	-0.004686
S.E. of δ	(0.00348)	(0.00393)	(0.00393)	(0.00402)	(0.00402)	(0.00463)
Prior-Year Test Score	Y	Y	Y	Y	Y	Y
Student Demographics		Y	Y	Y	Y	Y
Grade and Year F.E.s			Y	Y	Y	Y
School F.E.s				Y	Y	Y
Classroom Characteristics					Y	Y
Teacher F.E.s						Y
number of observations	753,912	536,960	536,960	536,960	536,960	536,960
R-squared	0.6984	0.7166	0.7167	0.7214	0.7215	0.7361

Note: $\delta 2$ is an estimate of the effect of repeating with a teacher on achievement when the student repeats with a teacher for a second time only. Where as, δ is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third, or fourth time. Lagged test scores consist of cubic-polynomials of math and reading scores. Classroom characteristics include number of students and cubics of lagged average classroom achievement in math and reading. *, **, *** represents significance at the 90, 95, and 99% levels, respectively

Table 6

Math Scores-Third, Fourth and Fifth Grade Sample

Regression	(1)	(2)	(3)	(4)	(5)	(6)
Estimate of δ_2	0.04003***	0.01568***	0.01026***	0.01415***	0.01425***	-0.00050
S.E. of δ_2	(0.00330)	(0.00370)	(.00369)	(.00372)	(.00372)	(0.00404)
Estimate of δ	0.03779***	0.01472***	0.009423***	0.01366***	0.01350***	-0.00037
S.E. of δ	(0.00326)	(0.00364)	(0.00364)	(0.00368)	(0.00367)	(0.00403)
Prior-Year Test Score	Y	Y	Y	Y	Y	Y
Student Demographics		Y	Y	Y	Y	Y
Grade and Year F.E.s			Y	Y	Y	Y
School F.E.s				Y	Y	Y
Classroom Characteristics					Y	Y
Teacher F.E.s						Y
number of observations	926,354	694,776	694,776	694,776	694,776	694,776
R-squared	0.6920	0.7145	0.7155	0.7249	0.7256	0.7604

Note: $\delta 2$ is an estimate of the effect of repeating with a teacher on achievement when the student repeats with a teacher for a second time only. Where as, δ is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third or fourth time. Lagged test scores consist of cubic-polynomials of math and reading scores that include third grade pre-test scores. The latter serves as second grade test score proxies. Due to missing data restrictions, current year attendance is used instead of lagged attendance. Classroom characteristics include number of students and cubics of lagged average classroom achievement in math and reading. *, **, *** represents significance at the 90, 95, and 99% levels, respectively

Table 7

Reading Scores-Third, Fourth and Fifth Grade Sample

Regression	(1)	(2)	(3)	(4)	(5)	(6)
Estimate of δ_2	0.03187***	0.006631*	0.004358	0.004829	0.005053	-0.006169
S.E. of δ_2	(0.00338)	(0.00379)	(0.00380)	(0.00386)	(0.00386)	(0.00435)
Estimate of δ	0.03098***	0.007176*	0.004917	0.004762	0.004854	-0.006308
S.E. of δ	(0.00333)	(0.00374)	(0.00374)	(0.00382)	(0.00382)	(0.00433)
Lagged Test Scores	Y	Y	Y	Y	Y	Y
Student Demographics		Y	Y	Y	Y	Y
Grade and Year F.E.s			Y	Y	Y	Y
School F.E.s				Y	Y	Y
Classroom Characteristics					Y	Y
Teacher F.E.s						Y
number of observations	923,871	692,892	692,892	692,892	692,892	692,892
R-squared	0.6797	0.6994	0.6998	0.7043	0.7045	0.7228

Note: $\delta 2$ is an estimate of the effect of repeating with a teacher on achievement when the student repeats with a teacher for a second time only. Where as, δ is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third or fourth time. Lagged test scores consist of cubic-polynomials of math and reading scores that include third grade pre-test scores. The latter serves as second grade test score proxies. Due to missing data restrictions, current year attendance is used instead of lagged attendance. Classroom characteristics include number of students and cubics of lagged average classroom achievement in math and reading. *, **, *** represents significance at the 90, 95, and 99% levels, respectively mates of δ_2 and δ from the math samples are no longer significantly different from zero. In the fourth and fifth grade math sample, the δ estimates reduce by 70%-75% and in the third through fifth grade sample estimates are reduced to less than zero.

A similar trend in the estimates of δ and δ_2 prevails in the reading samples as well. In both reading samples, estimates remain not significantly different from zero, however, they too are reduced to below zero. Magnitudes of the decrease in the reading estimates are 0.011-0.012 standard deviations of the test score distribution when moving from regression (5) to regression (6). This is similar to the reduction in the δ estimates from the math samples (0.012-0.014 standard deviations). In both math and reading, a large proportion of any possible positive effects of a student repeating with a teacher are accounted by teacher quality.

The large reduction in the value of the estimates of δ_2 and δ when teacher quality is controlled for gives rise to the possibility of omitted variable bias associated with higher quality teachers being more likely to repeat with students. I check for this nonrandom teacher assignment by testing for any classroom sorting based on whether students are more likely to repeat with a high-quality teacher than a low-quality one in the next section.

Nonrandom Classroom Assignment

Testing for High-Quality Teacher Repetition

Nonrandom classroom sorting is one explanation for why estimates may contain bias when critical variables, such as teacher fixed effects, are omitted from the baseline regression. When teacher quality is positively correlated with the likelihood of repeating with a student, then any estimation of the treatment effect that does not control for teacher quality will be upwardly biased. Positive correlation between the likelihood of repeating with a teacher and teacher quality comes from two sources.

Rothstein (2010) and Koedel and Betts (2011) find evidence of nonrandom classroom sorting when students' parents attempt to influence assignment to a teacher's classroom[21]. Presumably, students that have had a positive experience with their respective counterpart will rationally desire to pair with them again when given the opportunity. Specifically, if the expected utility of repeating with a teacher is greater than the expected utility gained of a random draw from the pool of other teachers in the grade, then the student's parents will rationally attempt to influence classroom assignment. Moreover, if the opposite is true and the expected utility of repeating with a counterpart is less than the expected utility of a random draw, the student's parents will not try to affect classroom assignment and may even attempt to avoid a repeat assignment. Both scenarios lead to a positive correlation between teacher quality and the probability of repeating in the same classroom again.

The second source of correlation comes from high-quality teachers having a greater probability than low-quality of teaching more than one grade. (B. A. Jacob & Lefgren, 2008) finds that principals can distinguish between teachers that are identified in the up- per and lower tails of the distribution of value-added measures. When principals use this identification ability to systematically assign higher-quality teachers to alternative teaching assignments, then, on average, higher-quality teachers will be more likely to repeat with students.

Both nonrandom allocation of teachers to classrooms/grades and the possible parental influence explained above create a positive relationship between teacher quality and the likelihood of teacher-student repetition. Regardless of the source of this correlation, determining whether there is an increase in the possibility of repeating with a student given a teacher's quality is imperative to validating that uncontrolled teacher quality is a source of bias in estimates of the δs in regressions (1) through (5) .

In examining the mean value-added estimates of teacher quality in Tables 1 through 3, one can deduce that average teacher quality for both subjects is substantially higher in both treatment groups, $D_2 = 1$ and $D = 1$. This rudimentary assessment is useful, but a more robust model is needed to validate any teacher sorting. In order to identify any nonrandom assignment, I test whether teacher quality increases the probability of a student repeating with a teacher by using the following model:

$$prob(D_{2it} = 1) = \sigma_{it} + \pi_t + \gamma_t + \pi_t * \gamma_t + \rho_i + \sum_{j=1}^{5} \phi_j Q_j + \nu_{it} \tag{7}$$

where σ_{it} represents student controls, π_t are time fixed effects, γ_t are grade fixed effects, ρ_i are school fixed effects, and the variables of interest, Q_j, are indicator variables representing the quintile of quality of the student's teacher. Using the Value-Added estimates from the baseline model (equation 6), teacher quality was found for math and reading by comparing teachers within the same school, grade, and year. From these estimates of teacher quality, in each subject teachers were assigned to quintiles, where Q_1 represents the bottom 20% of teachers, Q_2 represents the next 20%, and so on up until Q_5, the top 20% of teachers. When estimating (7), if there is a positive correlation between the likelihood of teacher repetition and quality of the teacher, this will be revealed by increasingly positive estimates on Q_j as j increases.

Sorting Results

Equation 7 is tested via a Limited Probability Model where the dependent variable is either $D_2 = 1$ for students that are repeating with a teacher and $D_2 = 0$ for students that are not. Results are reported in Tables 8 and 9 highlight that as teacher quality improves, students have a greater probability of having the teacher for a second time. When compared to teachers in the lowest quintile ($Q1 = 1$), in both math and reading, estimates on the indicator variable representing the top two teacher-quality quintiles ($Q4 = 1$ and $Q5 = 1$) are positive and significantly different from zero. Serial correlation of the errors at the classroom level is accounted for using clustered standard errors.

In the math samples, the average student has between a 3.11% (grades 3 through 5 sample) and 3.47% (grades 4 and 5 sample) chance of repeating with a low-quality math teacher ($Q1 = 1$). These probabilities increase, however, as the quintiles representing a teacher's quality increase. While the coefficient on the quintile representing teachers in the 21st-40th percentiles of teacher quality is not significantly different from zero, there is a significant increase in the probability of repeating with a teacher when the teacher is in the upper 60% of teacher quality. Specifically, when teacher quality is in the 61st- 80th percentiles and the 81st-100th percentiles, the probability of repeating with a teacher increases by 23%-26% and 29%-34%, respectively, when compared to a $Q1$ teacher.

In the grades 4 and 5 reading sample (Table 8), students have an unconditional probability of repeating with a teacher of 3.46%. This probability increases by 26% and 42% when the student has a reading teacher of quality $Q4 = 1$ and $Q5 = 1$, respectively. Increases in student-teacher repetition probabilities are even greater in the grades 3 through 5 reading sample. When compared to a $Q1 = 1$ quality teacher, all other teacher-quality types significantly raise the probability of repeating with a teacher. $Q2 = 1$ quality teachers in- crease the probability of teacher repetition by 28%, while $Q5 = 1$ quality teachers generate a striking increase of 46%.

The coefficients on a student's lagged achievement and economic background are included in Tables 8 and 9 to highlight the family-oriented educational inputs that affect the probability of students repeating with their teacher. The significance of the probability estimates on these variables suggest that students with greater family inputs are more likely to repeat with a teacher. Quantitatively, students from an economically disadvantaged household are 9%-10% less likely to repeat with a teacher.

Probit and Logit Limited Dependent Variable models with clustered standard errors at the classroom level were tested as robustness checks to the findings in Tables 8 and 9.[22] Across subjects and across grades, both probit and logit models maintain significance at the 95% level on the estimates of the top teacher quality quintile when compared to the bottom teacher quality quintile. Probability estimates in the grades 3 through 5 reading sample are again greater than in the grades 4 and 5 reading sample with estimates on the top two teacher quality quintiles continuing to be significant at the 95% level.

Discussion

As affirmed in the previous section, controlling for teacher quality is necessary when testing for any causal student achievement gain related to repeating with a teacher. Fortunately, the Value-Added Models framework can be implemented to control for these inputs. Using the VAMs' estimates of teacher quality, this research was able to test and evaluate the amount of classroom sorting that, when not controlled for, can lead to a significant misinterpretation of the estimated results.

Regressions (1) through (5) in Tables 4 and 6 highlight, on average, that a student who repeats with a teacher scores higher on standardized math tests than if that same student had not repeated with the same teacher. Results for reading regressions (1) through (5), although not always significantly different from zero, also produce positive estimates. These results agree with previous descriptive results (Cistone & Shneyderman, 2004; Rodriguez & Arenz, 2007) and initially seem to provide evidence about the merits of academic gains associated with looping. However, once measures are added to the model that control for differences in the abilities of the teacher (in (6) of Tables 4 through 7), gains in student achievement scores are similar between a student that repeats with a teacher and one that does not.

Table 8

LPM Estimates–Fourth and Fifth Grade Sample

Variable	**Math**	**Reading**
Teacher Q_2	0.003937	0.003162
	(0.00258)	(0.00292)
Teacher Q_3	0.006366**	-0.001862
	(0.00278)	(0.00305)
Teacher Q_4	0.006364**	0.007031**
	(0.00308)	(0.00334)
Teacher Q_5	0.008167**	0.01249***
	(0.00360)	(0.00394)
Lag Achievement	0.003003***	0.002284***
	(0.000380)	(0.000347)
Free/Reduced Lunch	-0.003064***	-0.003387***
	(0.000605)	(0.000605)
Other Student Controls	Y	Y
Grade and Year Fixed Effects	Y	Y
School Fixed Effects	Y	Y
Number of observations	538,063	534,253

LPM regressions include year, grade, and year by grade fixed effects. In addition to controls for lagged achievement and economically disadvantaged background a control for whether the student is repeating the grade is also included. Each teacher quality quintile is determined by ranking teachers based on their Value- Added estimates in math or reading. Standard errors are clustered at the teacher level. *, **, *** represents significance at the 90, 95, and 99% levels, respectively.

Table 9

LPM Estimates–Third, Fourth and Fifth Grade Sample

Variable	**Math**	**Reading**
Teacher Q_2	0.002476	0.006164***
	(0.00197)	(0.00201)
Teacher Q_3	0.003943*	0.008988****
	(0.00211)	(0.00217)
Teacher Q_4	0.007117***	0.009680***
	(0.00234)	(0.00229)
Teacher Q_5	0.009224***	0.009986***
	(0.00260)	(0.00259)
Lag Achievement	0.002556***	0.001838***
	(0.000315)	(0.00284)
Free/Reduced Lunch	-0.002370***	-0.002718***
	(0.000512)	(0.000505)
Other Student Controls	Y	Y
Grade and Year Fixed Effects	Y	Y
School Fixed Effects	Y	Y
Number of observations	698,777	701,634

LPM regressions include year, grade, and year by grade fixed effects. In addition to controls for lagged achievement and economically disadvantaged background a control for whether the student is repeating the grade is also included. Each teacher quality quintile is determined by ranking teachers based on their Value- Added estimates in math or reading. Standard errors are clustered at the teacher level. *, **, *** represents significance at the 90, 95, and 99% levels, respectively.

In both math and reading, δ estimates from the grades 3 through 5 sample are consistently lower than those from the grades 4 and 5 sample. Interestingly, the average achievement gain per student is greater when third graders are included in the treatment group. In treatment groups using the 4th and 5th grade samples (Tables 2 and 3), averaged normalized lagged test scores are slightly greater than current-year normalized test scores. However, when third graders are included in the treatment groups, average lagged test scores are lower than current-year test scores. This again allows for the average unconditional achievement growth to be greater when third graders are included in the sample, making the differences in the δ estimates and the fact that they continue to not be significant when teacher quality is controlled for a surprising result.

One explanation to the differences in the δ estimates has to do with differences in characteristics of students that repeat with teachers and those that do not. Students in the treatment group are more likely to come from an economically-disadvantaged household, be academically gifted, have a learning condition, be a limited English proficiency student and/or be repeating a grade. While controlling for these student characteristics allows for comparison between treated and untreated groups, it does not allow for a comparison of how the effect of repeating with a teacher differs for different types of students.

The possibility of heterogeneity in δ is evident in a) the differences of the δ estimates when 3rd graders are included in the sample and in b) the drastic reduction of the δ estimates when moving from regressions (1) to (2) in the grades 3 through 5 math and reading samples. Heterogeneity in δ may be based on grade as well as specific student characteristics and therefor may be driving the differences in the estimates of the δs across samples. For instance, if 3rd graders that repeat with a teacher are more likely to be academically gifted and the positive effect of repeating with a teacher is greater for academically gifted students, this would explain the differences in the δs when third graders are included and the drastic reduction of the δ estimates when moving from regressions (1) to (2) in the grades 3 through 5 math and reading samples.

The differences in the significance and the values of the estimates of the δs between the math and reading samples are not surprising and should be interpreted separately. For math, the estimates of δ are consistently less positive than the estimates of $\delta 2$ in regressions (1) through (5). Because observations where $\delta 2 = 1$ are a subset of observations where δ = 1, this finding indicates that any positive effects of repeating with a teacher are strongest in the first year the student and teacher repeat together (the second year together overall). This possibility of diminishing returns in the effects of repeating with a teacher warrants further investigation. Interestingly, in the math samples, when school fixed effects are added to the regressions in column (4), estimates of the δs increase. This may be associated with an increased likelihood of repeating with a teacher in certain schools or districts. Specifically, when schools with lower achievement gains on math tests are more likely to have students repeat with teachers, then estimates of the δs will increase when school fixed effects are added to the model. Schools that have systematically lower-achieving students have been known to be more likely to implement looping programs, specifically in urban areas in the U.S. This is because a greater proportion of students come from an insecure family structure in lower-performing schools than in high-performing schools. Further analysis of the heterogeneity of the δ estimates and their relationship to household inputs will help validate whether lower-performing schools should be more inclined to implement a looping policy. In the reading samples, regardless of the magnitude of the estimates, the reduction in the estimates of the δs is very similar to that of the math samples'. When going from regression (5) to (6), reading estimates reduce by similar amounts as the math estimates. This result highlights that the omitted variable bias associated with not controlling for teacher quality is similar across subjects. Estimates from the sorting models reaffirm this conjecture.

The sorting models featured in Tables 8 and 9 emphasize the degree of classroom sorting based on teacher quality. Notably, classroom sorting is greater in reading than math. Evidence of this is given via the larger reading estimates on the $Q5 = 1$ variables in the grades 4 and 5 sample. Furthermore, in the 3rd through 5th grade samples, an increased number of reading teacher-quality quintiles are significant when compared to math teacher-quality quintiles. While the estimates on the reading $Q5 = 1$ are about the same across subjects, estimates on $Qi = 1$ where $i = 2, 3, 4, 5$ are all significant when compared to a teacher quality of $Q1 = 1$. This is due either to low-quality teacher avoidance by the students (rather than high-quality teacher preference) or to principals being unlikely to assign low-quality teachers to a new grade.

Explanations as to why/how students sort to classrooms of higher quality teachers are related to one of two reasons or a combination of both: 1) When given the opportunity, students are more likely to lobby for high-quality teachers or 2) high-quality teachers are more likely to switch grades. An explanation of 1) is given in the previous section. Significance of estimates on family economic background and lagged achievement in the sorting models indicate that students with greater amounts of household educational inputs are more likely to repeat with a teacher than students with lower household inputs, presumably through heightened lobbying by parents.

An increased likelihood of high-quality teachers switching grades can be explained by a number of factors. There may be an incentive of an increase in pay or a bonus for teachers that switch grades. Teachers that are motivated to expand their teaching experience either because they want a new challenge or want to improve their resume will switch to a new grade. Teachers that switch grades may do so because they are dissatisfied with their current grade assignment and are looking for a better match between their skill set and their grade assignment. Lastly, mandatory grade changes may be implemented through specific school administrative policies.

Increases in pay and expansions of teaching experience create an incentive for high-quality teacher to switch grades at a higher rate than low-quality ones. When an increase in pay (or the prospect of an increase in pay from having extensive teaching experience in multiple grades) is the motivating device to switch grades, high-quality teachers are more likely to be accepted into the new role because of their more advanced skill set. This type of sorting exemplifies how, without any parental lobbying, additional administrative policies, or specific actions by the principal, students may organically have a higher chance of repeating with a high-quality teacher than a low-quality teacher.

The quality of teachers that migrate between grades because they are looking for a better fit or because of administrative policies is ambiguous. Depending on the administrative policy, principals may allocate lower-quality teachers to new grades in hopes of finding them a more productive role. However, principals may find it in their best interest to move higher-quality teachers to new grades in order to balance the teaching quality between grades. Regardless of these ambiguities and *why* students are more likely to repeat with high-quality teachers, this research emphasizes the importance of controlling for teacher quality when estimating treatment effects that are correlated with teacher assignment. Fortunately, VAMs provide the required framework to do so.

As with most research using VAMs, one concern when testing for the true effect on achievement is that students are non-randomly sorted between the treatment and control groups creating bias in the estimated results. Rothstein (2010), using data from the NCERDC, finds that 5th grade teachers have a large "effect" on 4th grade achievement [23]; in part, because of mean-reversion bias. He finds that a student's current classroom assignment is influenced by how much the student exceeded or fell short of expectations during the previous year. Fortunately, Rothstein (2009), shows that a large proportion of the bias associated with not controlling for non-random sorting can be mitigated by including a vector of student covariate controls. I include an extensive number of student-level variables in the baseline regressions to account for demographic differences between students in the treatment and control groups. The inclusion of specific student-level variables is largely modeled after the VAM specified in Chetty et al. (2014b) where they find very little bias in their estimates.

Perhaps the biggest criticism of the current investigation, is that the treatment variable does not really measure students that are looped with a teacher. Rather, this research answers a more defined question about repeating with a teacher. The lack of a positive effect from repeating with a teacher in regression (6) may be due to teachers having to pay an adjustment cost associated with teaching a new grade for the first time. When adjustment costs are high enough, any gain in student achievement from repeating with the teacher will be offset when the adjustment cost is passed on to the student.

Furthermore, when a teacher is looped with students, both the teacher and her pupils are aware that they will be spending (at least) another year together. Within-class teaching practices may be organized to maximize achievement based on this expectation. Teachers may have to spend less time preparing and organizing their class at the beginning and end of the school year allowing this time to be reallocated towards student achievement. Students may have better anticipation of a teacher's expectations heading into a new school year leading them to adjust their summertime knowledge accumulation during the time between looped school years. These influences are not measured in the treatment effect from the baseline regressions and, if large enough, could lead to a positive increase in student achievement from looping with a teacher.

This research has illuminated the upward bias in estimates measuring student achievement gains when teacher quality is not controlled for. When higher quality teachers are more likely to loop with students, positive estimated effects on achievement may be inappropriately associated with looping rather than the increase in teacher quality. Despite not specifically measuring the effects of looping in this investigation, the use of the VAMs has allowed for a causal interpretation of repeating with a teacher for longer than the normal school year. This technique can be extended to measure any gains to students that have been looped when compared to those that have not.

The use of VAMs, while controversial when used to determine teacher pay based on ranking of quality, is still very valuable in educational research. This research has employed its use as a control rather than a measure for ranking teachers; an application that has been used previously, (Jackson, 2009) and (Chetty, Hendren, & Katz, 2016). It is encouraged that the use of VAMs be implemented in other educational research investigations where it is necessary to control for teach quality.

Conclusion

While previous research has appropriately compared achievement gains of looped students with achievement gains of similar students that are not looped, necessary controls for teacher quality were missing. Given the strong positive relationship with teacher quality and the probability of repeating with a teacher, any inference of the estimated results needs to be made while holding teacher quality constant. This research employs these controls and ideally will serve as a building block for future, similar comparisons.

The comparison between students that repeat with a teacher to students that do not is a beginning step in determining the efficient amount of time a student and teacher should continue their educational relationship. While this investigation has asserted that academic gains, on average, are minimal when the student-teacher relationship is extended past 180 days, a few related questions are still unanswered.

Are there gains in non-achievement measures, such increased attendance or increased effort on homework assignments, that students experience from repeating with a teacher? Are there specific students that causally benefit from repeating with a teacher for a second year? Specifically, do students with unstable lives outside of school experience a greater benefit from a stable, long-lasting relationship with their teacher?

Despite the initial hypothesis that students would, on average, experience achievement gains from repeating with a teacher for another school year being rejected, this rejection does not rule out that a certain contingency of students would benefit from repeating with a teacher. Understanding any changes in the estimated effect in achievement of repeating with a teacher is crucial in determining the outcomes of looping teachers and students.

CHAPTER 2

NON-ACHIEVEMENT GAINS

Introduction

Recent literature on student outcomes has evolved to include both test scores and variables that are based on gains in non-cognitive, character traits. There has been substantial evidence indicating that interventions in elementary school have lasting, beneficial effects on character skills. Additionally, character skills have been shown to predict later-life outcomes with the same, or greater strength as cognitive measures (Heckman, Humphries, & Kautz, 2014).

Cunha, Heckman, and Schennach (2010) show that cognitive and non-cognitive skills are self-productive and exhibit dynamic complementarity; skills developed at younger ages directly and indirectly affect the productivity of future skills. They find that cognitive skill development (usually measured via achievement or IQ tests) is most sensitive during the years before age 11. However, the authors also find that character skills foster cognitive development, with investment in character skills during early years having a higher return than later investment.

Because of the influence character skills have on human capital development, this research complements the investigation in Chapter 1 by testing for any gains in a student's character skills from repeating with a teacher for one or more years. I use partial-persistence Value-Added Models (VAMs) to test for changes in teacher-subjective achievement scores, student-anticipated achievement scores, attendance, homework effort, and other measures of how students use their time outside of school from repeating with a teacher. I find that both students and teachers perceive increases in academic achievement and that effort by students increases. These results, combined with those from Chapter 1, highlight the necessity of using multiple measures of student achievement to understand the many facets of which teachers influence students.

The rest of the chapter is organized in following manner: Section II reviews the relevant literature for this investigation's research question, section III outlines the framework of the models, section IV provides information about the dataset used for empirical testing, section V gives the results, section VI discusses and specifies future relevant work, and section VII concludes.

Literature Review

The shortcomings of using test scores as the sole measure of human capital accumulation have been well documented. Adaptive tests, such as those where the question a student receives depends on his answers to previous questions, and linear interpretations of interval scaling (such that the increase from 400 to 450 on the SAT is interpreted as the same improvement in knowledge as an increase from 700 to 750) can lead to inappropriate comparisons across students (B. Jacob & Rothstein, 2016). Kane and Staiger (2001), using data from the NCERDC, argues that school-level test scores are noisy estimators of school quality estimating that about 38 percent of the variance in 5th grade reading test scores comes from sampling variation or yearly, idiosyncratic shocks. Furthermore Kane, Staiger, Grissmer, and Ladd (2002) finds that this uncontrolled volatility stems from a relatively low number of students per grade level and from one-time factors that influence test outcomes that are not sensitive to sample size.

If the primary goal of education is to develop a student's productive capabilities for when she enters the labor market, then skills associated with increased productivity are more valuable to employers and employees than outcomes on tests that serve as a proxy for these skills. Achievement tests are designed to measure acquired knowledge (crystallized intelligence) and IQ tests measure the rate at which people learn (fluid intelligence) (Heckman et al., 2014), but both types of tests are mistakenly used as the only measures of a student's skill set. Productive capabilities are, however, strongly related to both acquired knowledge and effort and character traits (Heckman et al., 2014).

Character skills [24] influence performance on a task either directly or indirectly, through changes in effort, and have been shown to be strong predictors of educational attainment, labor market success, health, and criminality. [25] Character skills that are crucial to a student's economic and social success include perseverance ("grit"), self-control, trust, attentiveness, self-esteem and self-efficacy, resilience to adversity, openness to experience, humility, tolerance to diverse opinions, and the ability to engage productively in society (Heckman et al., 2014).

Kautz, Heckman, Diris, Ter Weel, and Borghans (2014) compares cognitive and non-cognitive skill development, establishing that achievement tests do not adequately capture character skills that are valued in the labor market. The authors find that character skills rival cognitive skills in predicting labor market success. They highlight that character skills development is a dynamic process, where earlier investments lay the foundation for future investments. Cunha et al. (2010) agree that early-childhood investments in character skill development provide greater returns than later investments, especially for students with disadvantages in parental and birth endowments. Given the importance of investments in character skill development during elementary school, the current research investigates gains in these measures for students that repeat with a teacher in grades 4 and 5.

To test for character skill improvements associated with repeating with a teacher, I use a model similar to those found in Jackson (2016) and Bacher-Hicks, Chin, Kane, and Staiger (2017) Jackson (2016) uses a weighted average of the behavioral outcome variables of a student's suspensions, attendance, course grades, and on-time grade progression. He finds that, as long as test scores and behavioral outcome variables do not reflect the same exact mix of student skills: 1) some teachers may develop productive labor market skills without raising test scores and a researcher can better predict long-run life outcomes, such as labor market success, using both test scores and behavioral outcomes than using just test scores. In testing for bias in teacher performance, Bacher-Hicks, Chin, Kane, and Staiger (2017) use the Value-Added Model (VAM) framework to find little bias of teacher performance in value-added estimates, using test scores, scores from classroom observations, and student survey outcomes as the dependent variables in their models.

Psychologists have arrived at a relatively well-accepted taxonomy of character skills, with the acronym OCEAN, which stands for: Openness to experience, Conscientious- ness, Extraversion, Agreeableness, and Neuroticism (Heckman et al., 2014). Of the "Big Five" character skills, conscientiousness–the tendency to be organized, responsible, and hardworking–is the most widely predictive across a variety of life outcomes. Job performance, wages, and criminality are all most strongly related to conscientiousness (Heckman et al., 2014).

Framework for Testing the Effects on Non-Achievement

The framework used in this chapter closely follows the extended Value-Added Mod- els (VAMs) proposed in Chapter 1. These models build on the VAMs of Sass et al. (2014) and Chetty et al. (2014b). These VAM models measure a student's knowledge accumulation by specifying a production function in which current and past time-varying inputs are combined with a student's time-invariant inputs. Typically, knowledge accumulation is measured using test scores in math and/or reading. In this investigation, the achievement production function's dependent variable is replaced with measures of behavioral outcome variables, Y_{it}. Unlike Jackson (2016), I do not use a weighted average of the behavioral outcome measures. Rather test each outcome variable individually. I do this to identify which character skill is most significantly impacted from having the same teacher again.

In this investigations, I use the outcome variables of attendance, time spent on homework, and time spent reading as proxies for Conscientiousness, measuring a student's levels of perseverance and self-control. I use a student's anticipated grade in math or reading as a measure of trust, self-esteem, and self-efficacy. These traits are commonly associated with the "Big Five's" personality factors Extraversion and Agreeableness (Heckman et al., 2014). Lastly, I use the teacher's subjective assessment of the student's educational progress throughout the year as a measure associated with both cognitive and non-cognitive developments. A student's time-varying inputs are given by the vector, $\boldsymbol{V_i(t)}$, while a student's time-invariant component is denoted by μ_i. The generalized production function at time t is

$$Yit = Yt\,[\boldsymbol{Vi(t)},\, \mu_i,\, \upsilon_{it}\,] \tag{8}$$

Any unaccounted for influences on a non-cognitive measure are contained in the error term, υ_{it}, and the subscript on $Y_t[.]$ allows for the impact of the educational inputs to vary through-out time.

$\boldsymbol{V_i(t)}$, includes inputs at the school, classroom, and individual levels. These inputs are largely similar to the educational inputs specified in Chapter 1, with a few exceptions. School- and classroom-level inputs do not differ much from those that are typically included in achievement VAMs. However, school and classroom effects, including the teacher's input and any effect from a student and teacher repeating together, are now measured via changes in non-cognitive outcomes. Individual-level inputs include those specified in Chapter 1 plus a student's math and reading standardized test scores. Lagged achievement is usually included in achievement VAMs, however, it is normally used as a proxy variable for all past educational inputs. Conversely, in the VAMs used here, achievement measures are included to control for student heterogeneity and omitted educational influences, such as family inputs and time-varying student effort.

Assuming education production process is linear and additively separable and explicitly identifying the treatment variable that indicates whether a student has been matched with her current teacher previously, D_{it} from $\boldsymbol{V}_{it}$, equation (8) can be written

$$Y_{it} = \sum_{h=1}^{t} [\boldsymbol{\theta}_{ht}\boldsymbol{Z}_{ih} + \omega_{ht}D_{ih}] + \varphi_t \mu_i + \upsilon_{it} \tag{9}$$

The non-cognitive function above is time-specific, allowing the marginal effects of the independent variables to vary at each grade-time combination. These are found as average effects across students in the same school and grade during the same year. Equation (9) specifies that non-cognitive knowledge accumulation at time t is determined by the accumulation of all past and present, time-varying educational inputs; a student heterogeneity component, $\varphi_t \mu_i$, that can vary with t; and any idiosyncratic student-specific shocks. Usually VAMs are used to measure the influence of teachers on the outcome variable, however, in this research the VAM framework controls for confounding variables that may bias the parameter of interest, ω. ω measures the average treatment effect of having a teacher for a previous year on Y_{it}

To measure every coefficient found in equation (9) a researcher would need information on the full history of a student's educational inputs. To make estimation more tractable, two further assumptions are imposed. First, I assume the production function does not vary by grade level. Rather, the impact of an input varies only with the time span between the application of the input and the measurement of the dependent variable, h. This assumption is not overly restrictive because I only estimate education production functions across, at most, grades 3 through 5. Even if there are differences in the production of knowledge across grades, these differences should be mitigated due to the relatively small span of grades.

The second assumption states that the marginal impacts of prior educational inputs decay at the rate, λ, where $0 \leq \lambda < 1$ for any time span of application, h, such that $\boldsymbol{\theta}_{t-h} =$

$\lambda\boldsymbol{\theta}_{t-1-h}$, and $\omega_{t-h} = \lambda\omega_{t-1-h}$. This allows equation (9) to be written as:

$$Y_{it} = \sum_{h=0}^{t} \lambda^h [\boldsymbol{\theta Z}_{i,t-h} + \omega D_{i,t-h}] + \varphi_t \mu_i + \upsilon_{it} \tag{10}$$

Using a decayed lag of the (10)'s dependent variable, where $\lambda Y_{i,t-1} = \lambda \left(\sum_{h=0}^{t} \lambda^h [\boldsymbol{\theta Z}_{i,t-1-h} + \omega D_{i,t-1-h}] + \varphi_{t-1}\mu_i + \upsilon_{i,t-1} \right)$, the second assumption allows $Y_{i,t-1}$ to serve as a proxy for the input of all past educational variables on the current-year outcome measure. Importantly, this eliminates the need for a full history of past inputs. After collecting and simplifying like terms[26] equation (10) becomes:

$$Y_{it} = \boldsymbol{\theta Z}_{it} + \lambda Y_{i,t-1} + \omega D_{it} + (\varphi_t - \lambda\varphi_{t-1})\mu_i + \iota_{it} \tag{11}$$

Where $\iota_{it} = \upsilon_{it} - \lambda\upsilon_{i,t-1}$. Equation (4) specifies that non-cognitive measures are a function of current-year educational inputs, including D_{it}; a decayed proxy of all past educational inputs; student-specific heterogeneity; and any autocorrelated, idiosyncratic shocks. Student heterogeneity is accounted for by the inclusion of student-specific characteristics, including math and reading achievement. Equation (11) serves as the baseline model for this research.

I implement the baseline model by estimating (11) using a number of different specifications. I separately run regressions where the decay rate of educational inputs is $0 < \lambda < 1$ and when $\lambda = 0$. In the former, a lag of the dependent variable is included in the regressors, while in the latter it is not. Furthermore, I control for student heterogeneity by using either a student's current achievement, $A_{i,t}$, or her lagged achievement, $A_{i,t-1}$. When student heterogeneity is completely controlled for $\varphi_t = \lambda\varphi_{t-1}$, dropping $(\varphi_t - \lambda\varphi_{t-1})\mu_i$ from equation (11). This eliminates the need for the traditional control for student heterogeneity, student-level fixed effects.

Data

This research uses a longitudinal dataset from the North Carolina Education Research Data Center (NCERDC) to test the effects on character skills of repeating with a teacher. The dataset contains yearly information on a student's non-cognitive measures and characteristics as they progress through the state school system of North Carolina, including information on their assigned teacher, classroom, and school. The dataset spans the school years 2007 to 2013 and is restricted to elementary school students in self-contained classrooms. [27].

Chapter 2's investigation supplements Chapter 1's question about the *achievement* gains associated with repeating with a teacher. Therefore, I use a similar approach, where a simple dummy variable, D_n, was constructed to identify whether a student has previously been in his current teacher's classroom. Current-year, student-teacher matches were cross-referenced with a student's previous teachers for up to the past 4 years. For students that are currently being taught by a teacher they've had previously, $D_n = 1$. For all other students, $D_n = 0$, indicating that these students are being taught by a teacher for the first time.

The original dataset used to find whether a student has repeated with a teacher spans grades 1 through 5, but, because non-cognitive measures are unavailable for students in grades below 3rd grade, the testable population begins at grade 3. To prevent mistakenly associating one teacher's effect on character traits for another's, I cap the sample at fifth grade. Students are more likely to switch classes and have different teachers for different subjects in grades after fifth grade. Finding a teacher's true effect would require the nearly impossible task of disentangling multiple teachers' effects on a student's non-cognitive outcomes [28]. Similar to Chapter 1, 2.40% of the testable population repeat with a teacher for a second time ($D_2 = 1$) and 2.63% of the population are in a teacher's classroom that has taught them before ($D = 1$)[29].

I use a number of regressions to estimate the effect of repeating with a teacher on each non-cognitive measure. Each regression uses a distinct sample based on available independent variables. however, every sample is limited in the following ways: 1) Observations with missing information on their classroom teacher or on the school they attended are excluded. 2) Observations in classrooms with less than 7 and more than 45 students are excluded. 3) Observations from charter schools are excluded. 4) Observations that are missing information on individual student characteristics[30] are excluded.

The availability of lags of student achievement in math and reading and the availability of lags of each dependent variable determine the samples used to estimate the effect of repeating with a teacher on each non-cognitive measure. Non-cognitive measures are broken up into two groups; those based on a student or teacher's subjective assessment of student achievement[31] and those based on a student's time allocation. Data on the dependent variables is provided from the NCERDC using an interval scales, but these variables are standardized ($\mu = 0, \sigma = 1$) to make estimation results more interpretable.

Lags of these non-cognitive measures are unavailable before grade 3. Therefore, regressions that include lags of the dependent variable are limited to students in grades 4 and 5 and those repeating grade 3. Using lags of math and reading achievement in the equations limits samples to fourth and fifth grade students and third graders where scores of tests administered at the beginning of third grade act as second grade achievement proxy variables[32]. Lastly, as a robustness check, I also use samples that include only students that attend school for more than two-thirds of the year to account for any outlying outcomes that are abnormally affected by extraneous events. See Table 10 for a summary of the outcome variables.

Table 10

Non-Standardized Values of Character Trait Outcome Variables

Variable Name	Mean	Std. Dev.	Min	Max
Days in School	160.82	21.64	0	180
Time Reading	1.61	1.07	0	4
(1/2 hour intervals)	($\approx 45 min/day$)			
Time on Homework	1.99	1.08	0	5
(2 hour intervals)	($\approx 25 min/day$)			
Student Anticipated Grade-Math	2.73	1.01	0	4
(0=F, 1=D, 2=C, 3=B, 4=A)				
Teacher Judgment-Math	2.91	0.85	1	4
Student Anticipated Grade-Reading	2.77	0.98	0	4
(0=F, 1=D, 2=C, 3=B, 4=A)				
Teacher Judgment-Reading	2.90	0.84	1	4

Results

In each of the regressions used to find the estimates in tables 11 through 14, cubic polynomials of lagged math and reading achievement and controls for a student's age, gender, ethnicity, family affluence, any learning disabilities, switching schools during the year, and repeating the grade are included. Furthermore, differences in a student's classroom- and school-level inputs are controlled for using average classroom achievement and school fixed effects, respectively. Time and grade fixed effects along with their interaction are also included in all reported regressions.

In each of the tables 11 through 14, estimates of ω_2 and ω represent the non-cognitive effects of having a teacher for a second year ($D2 = 1$) and of repeating with a teacher ($D = 1$). In columns (1) and (3) of each table, controls for the teacher's input are not included, while they are included in columns (2) and (4). This offers a comparison between the two estimates of the ωs allowing for any student sorting based on teachers' inputs to be more easily highlighted. In all regressions, standard errors are clustered at the school level.

I run a number of regressions for each dependent variable as robustness checks. These include variations in the reported regressions such that the sample includes only students that have attended school for 120 days or more, a lag of the dependent variable is *not* included, and current achievement is used instead of lagged achievement. A lag of the dependent variable is included to allow for maximum flexibility of the value of λ in equation (11). When a lag of the dependent variable is not included, the restriction $\lambda = 0$ is assumed, meaning the impact of any past educational inputs has completely decayed by the start of current school year. I use this restriction to test for any differences in the ω's when third graders are included in the sample. In the reported results, only fourth- and fifth-graders are included in the sample, because information on the prior-year's dependent variable is not available before grade 3 [33].

Lagged achievement is preferred to current achievement because of the difficulties associated with disentangling the separate effects a teacher has on the cognitive and non-cognitive skills in the same year. For instance, consider a teacher that has an above average effect on a student's homework effort, but an average effect (value-added estimate = 0) on a student's cognitive achievement when measured independently of homework effort. If homework effort impacts a student's measured achievement, than the teacher's positive effect on homework effort may be misidentified as a direct effect on student's cognitive abilities, rather than an increase in the student's non-cognitive homework effort. Using lagged achievement rather than current-year achievement helps to eliminate this problem, so long as lagged achievement is unrelated to current teacher quality [34].

Changes to Subjective Measures of Achievement

Tables 11 and 12 report that, regardless of whether teacher quality controls are included, estimated effects on both subjective teacher assessment scores and student anticipated grade scores in the subjects of math and reading are significantly different from zero. Estimates of ω and $\omega 2$ in both tables are significant at the 99% level, except for estimates on subjective Reading scores when teacher fixed effects are included (column (4) of Table 11). These estimates are still significant at the 95% level.

In Table 11, estimated effects of repeating with a teacher on teacher-based subjectivity scores in math and reading, similar to achievement scores[35], diminish when teacher controls are included. However, unlike the findings in Chapter 1, estimates on subjective math and reading scores remain significantly different from zero when teacher controls are included. I estimate that repeating with a teacher increases subjective math scores by about 0.038 standard deviations and subjective reading scores by about 0.018-0.019 standard deviations. Unconditional on teacher quality, columns (1) and (3) report even greater increases in math (0.054-0.055 standard deviations) and reading (0.032 standard deviations) teacher-subjective scores.

In Table 11, estimated effects of repeating with a teacher on teacher-based subjectivity scores in math and reading, similar to achievement scores[36], diminish when teacher controls are included. However, unlike the findings in Chapter 1, estimates on subjective math and reading scores remain significantly different from zero when teacher controls are included. I estimate that repeating with a teacher increases subjective math scores by about 0.038 standard deviations and subjective reading scores by about 0.018-0.019 standard deviations. Unconditional on teacher quality, columns (1) and (3) report even greater increases in math (0.054-0.055 standard deviations) and reading (0.032 standard deviations) teacher-subjective scores.

Standardizing the dependent variables allows for the comparing the estimates of ω and $\omega 2$ to the estimates of δ and $\delta 2$ from Chapter 1. Most of Chapter 1's comparable estimates on math and reading achievement are not significantly different from zero. However, when they are, such as the estimates on math scores when teacher quality is not controlled for, the estimates from Chapter 1 are much smaller than this Chapter's estimates of the ω's on subjective teacher scores. When comparing column (1) of Table 11 to column (5) of Chapter 1's Table 4, increases in a teacher's subjective scores associated with repeating with a teacher are 3.4-3.5 times larger than the analogous measure on math achievement.

Increases in student achievement associated with being taught by a high quality teacher[37] is a useful standard to compare the estimates of the effect of repeating with a teacher on subjective teacher scores. Having a high-quality teacher increases a student's math and reading test scores by 0.16-0.18 and 0.12 standard deviations, respectively. Using estimates found in columns (2) and (4), increases in subjective teacher assessments are, respectively, about 22% and 15% of the increase in math and reading achievement associated with having a high-quality teacher.

The results on anticipated grades in math and reading, a measure related to a student's self-reflective assessment of their level of trust, self-esteem, and self-efficacy (Heckman et al., 2014), found in Table 12 highlight that students report higher anticipated scores when being taught by the same teacher again. When teacher fixed effects are not included in estimation ((1) and (3) in Table 12) the estimated effects of repeating with a teacher are slightly greater in math than in reading. However, both math and reading, regressions that include teacher-quality controls yield increases in anticipated grades of about 0.027-0.028 standard deviations. Estimates of the ωs are smaller in magnitude when teacher-quality controls are included, amounting to about 16% and 23% of the increase in math and reading achievement associated with having a high-quality teacher, respectively.

Table 11

Teacher Subjective Scores in Math and Reading

	Math		Reading	
Regression	(1)	(2)	(3)	(4)
Estimate of ω_2	0.05444***	0.03799***	0.03176***	0.01847**
S.E. of ω_2	(0.00653)	(0.00718)	(0.00663)	(0.00732)
Estimate of ω	0.05455***	0.03823***	0.03232***	0.01857**
S.E. of ω	(0.00647)	(0.00717)	(0.00658)	(0.00731)
Teacher F.E.s	N	Y	N	Y
number of observations	564,451	564,451	563,686	563,686
R-squared	0.5900	0.6441	0.5971	0.6478

Notes: Subjective scores are standardized scores based on the teacher's judgment of the student's achievement that year. ω_2 is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second time only. Where as, ω is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third, or fourth time. Controls for individual and institutional educational inputs included student lagged test scores that consist of cubic-polynomials of math and reading scores; student covariate controls for age, gender, ethnic- ity, family affluence, any learning disabilities, switching schools during the year, and repeating the grade; lags of teacher judgment scores in math or reading, respectively; classroom characteristics such as number of students and cubics of lagged average classroom achievement in math and read- ing; school, grade, and time fixed effects, including an interaction between grade and time fixed effects. The inclusion of lags of each regression's dependent variable (teacher judgment scores) presumes that $0 < \lambda < 1$ from equation (11) above. Reported standard errors of the are clustered at the school level. *, **, *** represents significance at the 90, 95, and 99% levels, respectively.

Table 12

Student Anticipated Scores in Math and Reading

	Math		Reading	
Regression	(1)	(2)	(3)	(4)
Estimate of ω_2	0.04021***	0.02707***	0.03373***	0.02656***
S.E. of ω_2	(0.00695)	(0.00760)	(0.00727)	(0.00796)
Estimate of ω	0.04057***	0.02776***	0.03425***	0.02796***
S.E. of ω	(0.00695)	(0.00766)	(0.00724)	(0.00797)
Teacher F.E.s	N	Y	N	Y
number of observations	565,095	565,095	564,510	564,510
R-squared	0.5516	0.6151	0.5510	0.6141

Notes: Anticipated scores are standardized scores based on a student's judgment of their expected achievement that year. ω_2 is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second time only. Where as, ω is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third, or fourth time. Controls for individual and institutional educational inputs included student lagged test scores that consist of cubic-polynomials of math and reading scores; student covariate controls for age, gender, ethnic- ity, family affluence, any learning disabilities, switching schools during the year, and repeating the grade; lags of teacher judgment scores in math or reading, respectively; classroom characteristics such as number of students and cubics of lagged average classroom achievement in math and read- ing; school, grade, and time fixed effects, including an interaction between grade and time fixed effects. The inclusion of lags of each regression's dependent variable (student anticipated scores) presumes that $0 < \lambda < 1$ from equation (11) above. Reported standard errors of the are clustered at the school level. *, **, *** represents significance at the 90, 95, and 99% levels, respectively.

Changes to Students' Use of Time and Effort

While the results in Tables 11 and 12 focus on outcomes that are largely related to achievement scores, the results in Tables 13 and 14 relate to how students allocate their time and effort. The outcome variable attendance is measured by number of days spent in school while the outcomes of time spent on homework, and time spent reading are all standardized ($\mu = 0, \sigma = 1$). Time spent on homework and time spent reading were initially reported based on interval scales.

Looking at Table 13, students attend about 2 more days of school when they repeat with the same teacher. Unconditional on teacher-based inputs, estimates of the ωs are 1.7-1.8 days of additional attendance, while, interestingly, the same estimates increase by about 25% when conditioned on teacher inputs. This is in direct contrast to the results found in Chapter 1. I examine the possible causes of this increase in the estimates of the ωs in (2) of Table 13 in the discussion section.

The estimates of the outcome variables in Table 14 helps us to understand if a student's repeating with the same teacher leads to changes in the student's allocation of time and effort *after* school. These outcome measures, like those in Tables 11 and 12, have been standardized to allow for comparison across variables. While time spent on homework best quantifies the amount of effort exerted by the student outside of the school setting, understanding whether any increases in homework effort come at the opportunity cost of other potential education inputs warrants investigation.

Estimates in column (1) of Table 14, indicate there is a positive and significantly different from zero increase in the amount of time a student spends on homework when they repeat with the same teacher, regardless of teacher quality. In columns (3) and (4) of Table 14, the negative coefficients on the ωs emphasize that students are spending *less* time

Table 13
Student Attendance

	Attendance	
Regression	(1)	(2)
Estimate of ω_2	1.740***	2.146***
S.E. of ω_2	(0.101)	(0.117)
Estimate of ω	1.799***	2.262***
S.E. of ω	(0.102)	(0.119)
Teacher F.E.s	N	Y
number of observations	569,838	569,838
R-squared	0.2499	0.2934

Notes: Attendance values and their associated estimates are measured in number of days. ω_2 is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second time only. Where as, ω is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third, or fourth time. Controls for individual and institutional educational inputs included student lagged test scores that consist of cubic-polynomials of math and reading scores; student covariate controls for age, gender, ethnicity, family affluence, any learning disabilities, switching schools during the year, and repeating the grade; lags of teacher judgment scores in math or reading, respectively; classroom characteristics such as number of students and cubics of lagged average classroom achievement in math and reading; school, grade, and time fixed effects, including an interaction between grade and time fixed effects. The inclusion of lags of attendance in the regressions (1) and (2) presumes that $0 < \lambda < 1$ from equation (11) above. Reported standard errors of the are clustered at the school level. *, **, *** represents significance at the 90, 95, and 99% levels, respectively.

Table 14
Time Spent on Homework and Reading

	Homework		Reading	
Regression	(1)	(2)	(3)	(4)
Estimate of ω_2	0.03384**	-0.00476	-0.01540*	-0.01657*
S.E. of ω_2	(0.0134)	(0.0147)	(0.00806)	(0.0103)
Estimate of ω	0.03344***	-0.00185	-0.01549*	-0.01623*
S.E. of ω	(0.0136)	(0.0148)	(0.00797)	(0.0102)
Teacher F.E.s	N	Y	N	Y
number of observations	422,305	422,305	426,319	426,319
R-squared	0.1039	0.2305	0.1759	0.2194

Notes: Time spent on homework and reading is standardized with mean = 0 and standard deviation= 1 to allow for comparison across dependent variables. ω_2 is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second time only. Where as, ω is an estimate of the effect of repeating with a teacher when the student repeats with a teacher for a second, third, or fourth time. Controls for individual and institutional educational inputs included student lagged test scores that consist of cubic-polynomials of math and reading scores; student covariate controls for age, gender, ethnicity, family affluence, any learning disabilities, switching schools during the year, and repeating the grade; lags of teacher judgment scores in math or reading, respectively; classroom characteristics such as number of students and cubics of lagged average classroom achievement in math and reading; school, grade, and time fixed effects, including an interaction between grade and time fixed effects. The inclusion of lags of each regression's dependent variable (time spent on homework in regressions (1) and (2) and time spent reading in (3) and (4)) presumes that $0 < \lambda < 1$ from equation (11) above. Reported standard errors of the are clustered at the school level. *, **, *** represents significance at the 90, 95, and 99% levels, respectively.

reading when they repeat with the same teacher. While the estimates of the increase in time spent on homework are more than twice as large as the estimated decrease on reading time, together these results highlight that students partially offset the increase in homework time with less time spent reading.

Specifically, the estimates on time spent on homework in column (1) show that students move up the distribution of homework time by 0.033-0.034 standard deviations. Treatment estimates of the original, non-standardized homework variable help us to understand this result. The original values of the homework variable are measured as an inconsistent interval scale, with outcomes 0-5 [38]. Estimates of the non-standardized variable[39] are not much different than those for the standardized outcome, with values of 0.035-0.036[40]. A student from the sample averages 1 to 3 hours per week doing homework, unconditional of any covariates. A back-of-the-envelope calculation, simplifying the average homework time to 2 hours per week, finds that a student spends a little more than 4 additional minutes doing homework per week when they repeat with a teacher.

While students that repeat with the same teacher spend more time doing homework, adding teacher Fixed Effects show that this increase is mostly related to the teachers' input of those that repeat with students. In (2) of Table 14, estimates of time on homework are no longer significantly different from zero, where, in (4), estimates on reading time are still negative and significantly different from zero. The R^2 of the Homework regression more than doubles when teacher inputs are included, highlighting the influence teachers have in determining the amount of time a student spends on homework. Including teacher inputs when measuring the effects of repeating with the same teacher on homework effort is essential to determine whether students are working harder because they are repeating with a familiar teacher or because the teacher they repeat with systematically assigns more homework. The findings in Table 14 indicate the latter.

Discussion

The benefits of repeating with a teacher for another school year appear to be stronger in non-cognitive measures than in the scores on objective achievement tests studied in Chapter 1. Character traits associated with a student's levels of Conscientiousness, Extraversion, and Agreeableness all see positive gains when she repeats with a teacher, providing evidence of the educational enhancements associated with programs designed to keep students and teachers together for multiple years. Like the regressions using test scores as outcome variables, one must control for nonrandom student sorting into teachers' classrooms when using non-cognitive measures. However, unlike the estimates of academic gains in math and reading test scores, the estimates on the gains in subjective teacher scores and anticipated student scores are significantly different from zero even after controlling for teacher quality. These results, in combination with findings supporting increased student effort when repeating with a teacher, provide substantial evidence for the character-building benefits of keeping students and teachers together for multiple school-years.

Rockoff and Speroni (2011) finds that subjective and objective measures of teacher quality are equally predictive of future student success. Furthermore, subjective scores of student success may be based on academic achievement success as well as character skill gains where objective subject test outcomes only measure the former. Therefore, in understanding the complete academic impact of repeating with a teacher, one needs to look at both objective and subjective outcomes. While Chapter 1 establishes that objective outcomes are non-significantly affected by repeating with a teacher, subjective outcomes see significantly positive gains. When subjective scores account for character skills as well as academic achievement measured by test scores, the significant effects on subjective scores and not objective scores emphasizes that a student's character skill development is more strongly affected than academic achievement when a teacher-student relationship is maintained.

If the increase in a teacher's subjective assessment of a student's academic success is due to nonrandom measurement bias (such as when a teacher's evaluations of students are affected by increases in the amount of time spent with them), then the results from subjective scores can be misleading. However, the gains in student perceptions of their academic success, as measured by anticipated scores in math and reading, complement the results from subjective scores. Both teachers and students perceive increased academic success when they repeat together, highlighting that both parties are more optimistic about their educational experiences. This optimism, categorized as an increase in the psychological character trait, extraversion (Heckman et al., 2014), fosters an environment more conducive to learning and helps establish a more positive student outlook that can lead to further academic and adult-life success.

Given the nonrandom student sorting into higher-quality teachers' classrooms established in Chapter 1, the dampened estimates when teacher input controls are included in the teacher-subjective and student-anticipated regressions is not surprising. However, the increase in the estimated effect of repeating with a teacher on student attendance when teacher controls are included in the regression (Table 13) is surprising, especially when coupled with the results from the regressions using student time spent on homework as the dependent variable. Both attendance and time spent on homework are measures of student effort (Heckman et al., 2014). When teacher controls are included in the regressions using time spent on homework as the dependent variable, estimates are dampened and no longer significantly different from zero.

These results suggest that, conditional on teacher quality, students are more willing to come to school when they are repeating with a teacher. However, this does not make students increase their out-of-school effort. This again suggests that, on average, repeating with a teacher makes school more enjoyable for the student. This improvement in a student's feelings about school is especially important for elementary school students. Cunha et al. (2010) and Heckman et al. (2014) both find evidence that investments in character skills during early childhood years have a higher economic return than investment in later years because they build the base for subsequent investments. This concept of *self-productivity* can be summarized as "skills beget skills" (Heckman et al., 2014).

Interestingly, the lack of significance on estimates measuring increases in homework effort when teacher controls are included and the previously established nonrandom sorting of students to higher-quality teachers suggest a causal relationship between homework effort and high-quality teachers. Students, on average and unconditional of teacher quality, increase their homework effort and see higher achievement when they repeat with a teacher. However, after controlling for teacher inputs, a student's homework effort and achievement gains are not significantly affected by repeating with a teacher. This suggests that higher-quality teachers increase a student's homework effort (either by assigning more homework or inspiring students to spend more time on their assignments) which, in turn, increases the student's test scores.

Taken together, the effects of repeating with a teacher favor a student's non-cognitive, character skill development more than the development of cognitive skills that are traditionally measured by test scores. This falls in line with previous research that finds teachers to have a bigger effect on character skills than on cognitive skills (Jackson, 2013b). Furthermore, these two teacher effects are only weakly correlated, suggesting that some teachers improve only one dimension of skill development without improving the other (Jackson, 2013b).

The current investigation has asserted that, on average, repeating with a teacher makes school more enjoyable for a student and students that enjoy school do better on standardized tests (Heckman et al., 2014). However, despite the positive effect repeating with a teacher has on a student's perception of school, contemporaneous student test scores do not see an improvement. I hypothesize this is because increases in a student's character skills occur more quickly than changes in a student's academic achievement. It seems that any gains in academic achievement from repeating with a teacher will be associated with increases in character skills and not from a student learning more of the course curriculum. These character skill gains are still developing during the year a student and teacher are repeating, with their full effect not being realized until the school-year is over and the standardized tests have been taken. An interesting extension of the current research would be to test for any gains in math and reading achievement in the year *after* a student repeats with a teacher. This research does not rule out all math and reading achievement gains from repeating with a teacher; only contemporaneous achievement gains. The findings in this investigation stress that programs intended to keep students and teachers together for multiple years should be implemented during elementary school rather than in high school because the benefits of character skill improvements are greatest among younger children. Furthermore, this research highlights the importance of using multiple dimensions when measuring gains in knowledge and productive capabilities.

Conclusion

This research has identified the improvements in non-cognitive abilities elementary school students experience when repeating with a teacher. The gains in character traits established through this investigation are important, because, when measuring academic success purely on the cognitive gains associated with achievement tests, a researcher will fail to recognize academic gains that will lead to greater amounts of school completion and better results in the labor market. Gains in character traits should be investigated as complement measures of research using test scores as a mark of academic success.

Both students and teachers perceive increases in the academic achievement when "looped" together. Subjective scoring (by teachers) and anticipated grades (by students) show gains when each stay together longer than the usual 180 days. Additionally, student effort improves, as measured by attendance. These findings establish that despite the lack of improvement in test scores when students repeating with a teacher, students are experiencing academic gains when they repeat with a teacher for another year.

Given the differences in non-cognitive and cognitive gains of repeating with a teacher, it is important for future work to identify any heterogeneity in these effects. Kautz et al. (2014) present evidence about the increases in character gains of students from disadvantaged backgrounds compared to students from stable backgrounds of programs implemented to improve the "soft skills" of students. This is an avenue for future work; not just to measure of changes character gains, but also to test for any heterogeneity in achievement gains.

CHAPTER 3

REVERSING BROKEN WINDOWS THEORY

Introduction

Since the 1970s, criminologists have shown that "fear of crime" is distinct from "fear" of "crime"—residents are fearful not only of actual crime, but of other physical and social cues of disorder and incivilities embedded within their environments (Garofalo & Laub, 1978). Research in this realm has continued to pique interest because incivility perceptions have been linked to a number of untoward outcomes. They can affect stress and well-being, and consequently health (Ross & Mirowsky, 2001); reduce participation in local community organizations (Perkins, Brown, & Taylor, 1996); decrease the likelihood that residents will walk outdoors in their own neighborhoods (Gallagher, et al., 2010); lead to increases in later crime (Taylor, 2005); and might even predict the decline of entire neighborhoods (Skogan, 1990). There is empirical support for the cross-sectional, individual level incivilities model (incivilities→fear). LaGrange, Ferraro, and Supancic (1992) and Wyant (2008) demonstrated that the causal chain is largely incivilities→risk perceptions→fear. However, these and other works have relied on cross-sectional data, which creates the potential problem of causal directionality.

Wilson and Kelling's (1982) "Broken Windows" version made the model ecological and longitudinal. Their version argued that physical and social incivilities clustered on street blocks can lead to reduced informal control, thereby creating a concentration of emboldened potential offenders. More fear and withdrawal occurs over time, and more crime-inclined people move in. Limited empirical support for this model exists. To date, two studies have examined the lagged impacts of incivilities on neighborhoods and street blocks (Robinson et al., 2003; Taylor, 2001). Taylor (2001) found that incivilities showed some lagged impacts on crime, decline, and fear; however, these impacts varied across outcomes and depended on the type of incivility indicator used. Further, neighborhood structural variables had comparatively larger impacts. Robinson et al. (2003) used a one-year follow-up design and found that earlier perceived incivilities show weaker-than-expected impacts on later changes in several reactions to crime—including fear. In short, longitudinal support for BWT is limited.

Recent work emerging from the incivilities paradigm emphasizes the subjective nature of disorder perceptions (Hipp, 2010; Sampson, 2009; Sampson & Raudenbush, 2004; Wallace, 2012; Wickes et al., 2013). This literature shows us that the assumption within BWT that disorder cues can be summarized by "we know it when we see it" rests on tenuous empirical grounds, and that perhaps Harcourt (2001) was correct to argue that a clear and rigid category of "disorderly" does not exist. In light of these findings, and in conjunction with BWT's modest level of empirical support, should the beginning sequence of the theory be reconfigured to reflect the variable nature of perception? Are there plausible theoretical reasons, for example, to reverse the initial link of BWT by arguing that risk perceptions may shape perceptions of incivilities? The current work—relying on a symbolic interactionist perspective on incivilities (Harcourt 2001)—argues for such a plausible relationship in that people feeling high localized crime risk might be "biased" (Hipp, 2010) toward interpreting their environments as more incivility-ridden. In other words, the person perceiving more crime risk might be more attuned to the immediate surrounding problems than those perceiving less. Similar to BWT, this model amounts to a psychological effect with an individual-level outcome. Given the longitudinal nature of our data, we can test both causal pathways (incivilities→risk and risk→incivilities).

Wilson and Kelling (1982) illustrate that some dynamics that affect individual-level outcomes occur at higher levels, such as the street block or the neighborhood. It is within these specific street block or neighborhood contexts where crime-inclined people can proliferate. Thus, the Wilson and Kelling argument amounts to a multilevel, ecological argument (street block-level crime and incivilities→more fear among group members). Accordingly, we also build an ecological test of BWT and its reverse into our models by aggregating risk and incivilities perceptions to the street block level and testing both pathways. We ask: Do earlier incivilities perceptions aggregated to the street block have impacts on later changes in individual perceptions of risk? And: Do earlier risk perceptions aggregated to the street block level have impacts on later changes in individual perceptions of street block problems? Which model finds more empirical support?

The current work takes advantage of longitudinal and multilevel data to test BWT and its reverse formulation with risk perceptions impacting how one sees her environment. This is the first study to examine this reverse ordering. In a closely related investigation with a slightly different outcome, Brunton-Smith (2011) used a cross-lagged design to examine the impacts of incivilities on fear and fear on incivilities and found more support for the traditional BWT model. Our findings diverge from his. While we do find limited support for the traditional BWT model, two different types of models show more support for the reversed model. HLM models and fixed effect analyses show substantial support for a model by which risk perceptions shape the degree to which people see things in their environs as problematic later in time.

Biased Perceptions

Incivilities thesis models and key empirical findings are outlined below. We then tie in an emerging literature that questions the basic starting point of the BWT model—that neighborhood incivilities, or embedded neighborhood cues, are universally perceived and defined as problematic. This literature investigates what factors shape or "bias" perceptions of neighborhood problems (Franzini et al., 2008; Harcourt, 2001; Hipp, 2010; Sampson & Raudenbush, 2004; Wallace, 2011; Wickes et al., 2013). With this literature's focus on the variability in how people define neighborhood problems as the backdrop, the current work exploits a symbolic interactionist perspective to elucidate how perceptions of incivilities may also be a function of feelings of localized crime risk. Our empirical findings not only question the theoretical adequacy of the current articulation of BWT, they highlight another reason that perceptions of local incivilities are not the unambiguous and universal cues they were once thought to be; rather, they are subjective constructions that are shaped by many biases, including crime risk perception.

Incivilities Thesis Research

Victimization surveys have revealed that fear of crime, especially in urban environments, is more ubiquitous than the actual number of people who became victims of crime (Cook & Skogan, 1984; Hunter, 1978; Taylor, 1999). This is why many have argued that "fear of crime" in urban environments is more complex than simply "fear" of "crime" (Garofalo & Laub, 1978; Hunter, 1978; Wilson, 1975). Linking social incivilities (i.e., loitering youth, panhandlers, and public drunkenness) and physical incivilities (i.e., litter, vandalism, vacant and/or rundown structures) (Perkins & Taylor, 1996), the initial research was individual-level and focused on how the presence of incivilities within urban communities contributes to increased levels of fear for community residents. Wilson and Kelling's (1982) version asserted that incivilities in certain neighborhoods caused law-abiding residents to avoid those places out of fear, while those inclined to anti-social behavior would remain. The persistence of these incivilities sends a signal to local teens and at-risk youth that the area is ripe for delinquency (Taylor & Covington, 1993). On these street blocks and in these neighborhoods, the remaining concentration of residents with proclivities toward crime would thus produce an increase in the frequency and severity of crime and other anti-social behaviors. As such, their model predicts that crime will increase in areas with high levels of incivilities over time through social psychological processes that occur at the street block and neighborhood levels.

Several years later, Skogan (1990) posited a version of the thesis (in "Decline and Disorder") whereby incivilities begin a process that leads to fear primarily through risk perceptions occurring at the *neighborhood* level. Fear of crime and weakened local commitment may, as Skogan argued, lead to later deterioration at the neighborhood level. Expanding on the Wilson and Kelling framework, Skogan argued that disorder not only might lead to later elevated crime, but to an increase in later incivilities as well. Testing his theory, Skogan analyzed cross-sectional data from five large American cities and found that high levels of incivilities predicted neighborhood decline as manifested by increased crime later in time.

Since 1975, research and theory on the incivilities thesis has gripped academics and public policymakers alike on a broad basis. However, empirical tests of the theory have been less than strongly supportive (Harcourt & Ludwig, 2006; Harcourt, 2001; Taylor, 2001). In 2001, Harcourt re-analyzed Skogan's (1990) data and found only a correlation between levels of disorder and later robbery, suggesting that his framework's ability to predict crime generally is limited. Taylor's (2001) work only partially supported the thesis when he found that minor physical decay and social incivilities in Baltimore were linked with only some types of serious crime changes, but not others. Further, Taylor found that the impacts of these incivilities paled in comparison to the structural variables in his models. More recently, and inspired from earlier writings (Taylor, 2001), Gau and Pratt (2008) critiqued BWT on a conceptual level by showing that "Citizens did not seem to differentiate between disorder and crime; rather, the two blended together in their eyes". If this is so, BWT may need revision on a fundamental level.

The vast majority of scholarship in the broken windows realm has argued theoretically for the impact of incivilities on risk and fear (most well known in Wilson & Kelling, 1982). However, most of this research has used cross-sectional data to support its assertions (Skogan, 1990; Wyant, 2008). This elevates the problem of causal directionality:

Is it perceiving incivilities that leads to risk perception, fear, and withdrawal, or is it people who see more local risk defining environmental cues as disorderly (Perkins & Taylor, 1996; Taylor, 1999)? Some critics have argued that relationship may be entirely spurious, as both perceptions of disorder and risk may be driven by unmeasured neighborhood processes (Gau & Pratt, 2008; Kubrin, 2008; Sampson & Raudenbush, 1999; Worrall, 2006; Xu et al., 2005).

What is "Disorder" and "Incivility"?

Another challenge in this line of research centers on the need for conceptual clarity regarding the nature of incivility and disorder (Kubrin, 2008). Challenging the assumption that the perception of incivilities is an objective assessment of physical indicators of disorder, researchers have found evidence that people define disorder differently. Akin to George Kelly's (1955) theory of personal constructs which finds that "the correspondence between what people really think exists and what really does exist is a continually changing one", the inclination to perceive a certain indicator as a mark of disorder or incivility is dependent on different variables at both the individual and group levels. These variables include race, social cohesion, neighborhood attachment, and exposure to teens hanging out, among other factors (Hipp, 2010; Wallace, 2011; Wickes et al., 2013).

Previously, nearly all prior work on BWT focused on how incivilities and other neighborhood characteristics shape risk and risk and fear outcomes—either at individual or neighborhood levels (Markowitz et al., 2001; McGarrell, Giacomazzi, & Thurman, 1997; Robinson et al., 2003) with little questioning of how incivilities are perceived. In contrast, recent research has investigated how perceptions of neighborhood problems are shaped (Carvalho & Lewis, 2003; Franzini et al., 2008; Sampson & Raudenbush, 2004). Much of this work has investigated how perceptions of incivilities may not be simple reactions to unambiguous symbols that are noticed by passersby in the same manner— as

"Broken Windows" theory may suggest (Harcourt, 2001). Rather, social and physical cues are defined and socially interpreted through a filtering process, causing more perceived problems among particular individuals, as well as among particular areas. The limited research in this area supports the idea that incivilities perceptions are not solely a function of actual incivilities, but rather they are A) socially constructed—at least in part (Hipp, 2010; Wallace, 2011), and B) shaped by neighborhood structure such as racial and class composition (Franzini et al., 2008; Jackson, 2004; Sampson & Raudenbush, 2004; Wickes et al., 2013). In a similar vein, Innes (2004; 2005) argued that it is not per se crime or incivilities rates in an area that necessarily shapes residents' perceptions, but it is how certain types of events—or "signals"—are processed and defined by members of the community. Exactly what makes an event a "signal" event, however, is not well understood.

The link between "objective" and community or individual assessments of incivility is not as clear as the traditional model of BWT indicates. One cause for concern is the difference in findings when objective indicators rather than survey respondent assessments of disorder are used. Studies incorporating the former, either via systematic observation of street blocks (Perkins & Taylor 1996; Sampson & Raudenbush, 2004; Taylor & Covington 1993; Taylor 2001), or interviewer assessments (Brunton-Smith & Sturgis 2011; Skogan 1990; Taub, Taylor, & Dunham 1984), found a weaker correlation between incivilities and fear than studies using subjective respondent measures. As Taylor (1999; 2001) and others have pointed out, these findings might indicate that incivilities and risk may be linked in ways we currently do not understand. In other words, cross-sectional associations found between incivilities perceptions and the perception of risk may represent, in part, a phenomenon by which people define situations as incivility-ridden based on perceived risk. But is there a plausible theoretical basis to argue that perception of crime risk would shape how one defines ambiguous social and environmental cues?

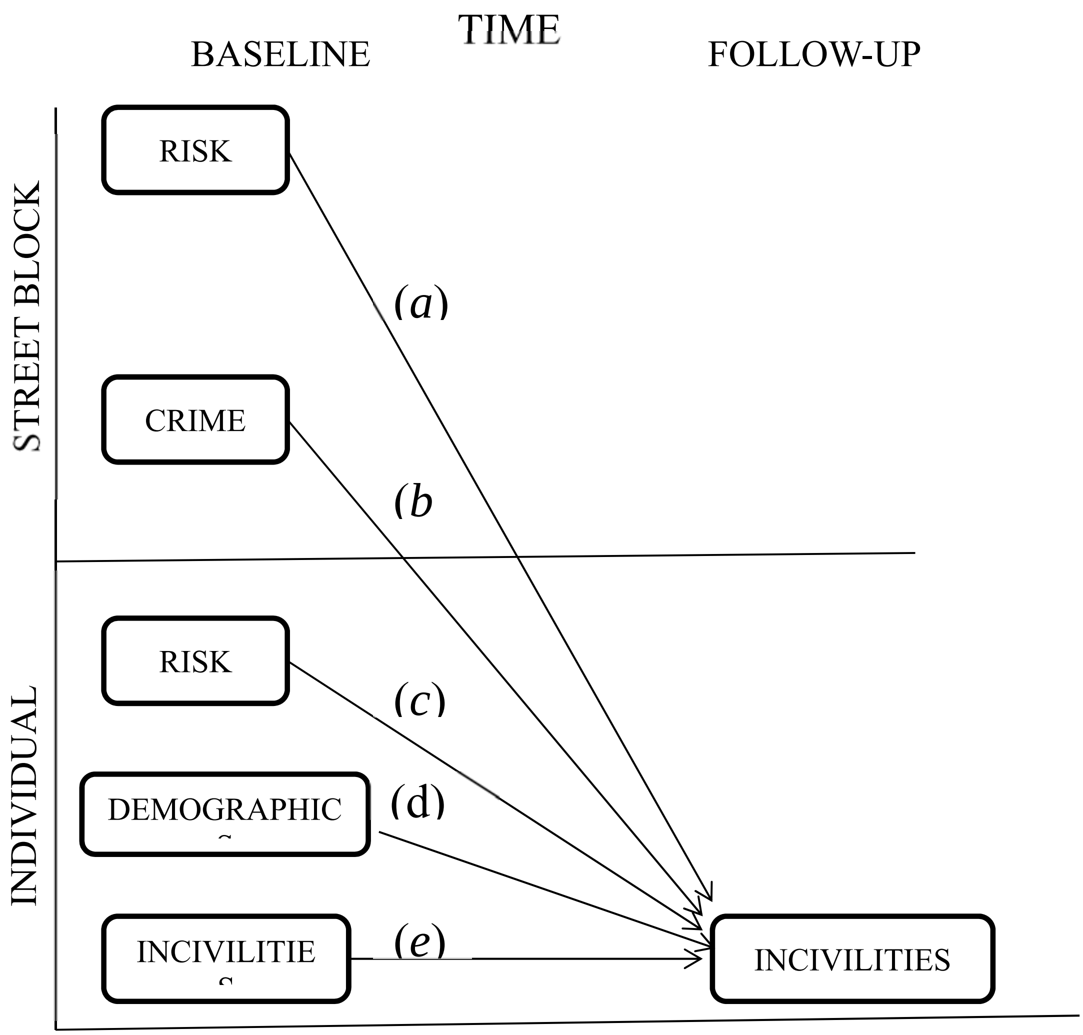
Current Investigation

We draw on Harcourt's (2001) framework to argue for a theoretical relationship by which perceived risk affects perceptions of incivilities. Harcourt contends that BWT and order maintenance policing is problematic because “the norm and meaning of orderliness have become fixed, natural, and uncontested”. Instead, Harcourt argues that context and social meaning are the driving force behind what is understood as disorder for a given individual or community. In essence, he argues that theoretical categories of “orderly” and “disorderly” are social constructions rather than reflections of empirical reality.

This phenomenological approach, in appreciating the subjective nature of human experience, emerged out of the symbolic interactionist perspective (Matsueda, 1992; Mead, 1934). Mead (1934) argued that humans do not perceive their social environment directly and universally; rather, they define and give personal meaning to situations they are in. In this sense, perceived realities and interactions are social constructions that will vary from person to person. How individuals define these situations is based on ongoing and evolving social interactions and thought processes. Applying this framework to the incivilities/risk association, a reasonable argument could be made that perceiving incivilities in one’s surroundings is a subjective and ongoing process that is shaped by changing interactions with humans within the same ecology, as well as changing cognitive assessments of danger. If this is true, individuals may assess their social environments as being more or less deteriorated based on the degree to which they see and interpret ongoing ambiguous cues (or “signals”) as fear-inducing.

Further, there is reason to believe that the impact of risk perceptions on perceived neighborhood problems might be stronger in some areas. Neighborhood stability can promote ties to the community (Bursik & Grasmick, 1993), and can make residents feel less worried about crime and other neighborhood issues (Taylor, 1996), and thus might

Figure 1
Conceptual Model



prevent residents from interpreting possibly ambiguous cues as disorderly. Also never tested before, our data allow us to test the hypothesis that people living on street blocks with higher levels of stability are “protected” from reacting to feelings of risk by defining neighborhood cues as problematic.

The above conceptual model illustrates how both models are tested (solid lines indicate BWT; broken lines indicate our reversed hypotheses). Using data from 50 street blocks in Baltimore City, we examine the impact of pathway *b* (broken line)—the impact that risk perceptions at the street block level at baseline has on later changes in perceptions of incivilities at follow-up, while controlling for baseline incivilities at the individual (pathway *j*) and other key controls. A parallel investigation on the individual level—pathway *f*—will examine the impact that risk at baseline has on later incivilities at follow-up, while holding baseline incivilities and other predictors constant. In an attempt to control for potential confounds, we also hold constant demographics at the individual level and recent crime changes at the street block level (pathways *d* & *h*) to see if the impacts of pathways *b* & *f* survive. Respectively, pathways *a* and *i* test the street block- and individual-level models as articulated in the traditional BWT model.

Methods

Setting

Panel data used in this analysis derived from a probability sample of Baltimore City residents (for a detailed description of the data and procedures, see Perkins 1989). Baltimore during this period was not unlike many other older urban cores in the U.S. that were experiencing the harsh effects of deindustrialization (Taylor, 2001). Once a booming industrial center featuring a bustling maritime port, the City experienced a 6.5% decrease in population and dramatic job losses between 1980-1990, marked by a 38% reduction in manufacturing jobs, a 12% drop in the trades, and a 21% decrease in infrastructure jobs.

(U.S. Bureau of Economic Analysis, 2011). Relatedly, the population was poor, with median household incomes of $12,811 and $24,045 in 1980 and 1990, respectively, and in the same years 56% and 33% of households earned incomes of less than $15,000, (in 2009 inflation-adjusted dollars). The population was young, with 60% between ages 5 and 44 in 1980, and 61% between the same ages in 1990. A city characterized by a sizable African American population, 56% of the population in 1980 was non-white, and this rose to 60% by 1990 (U.S. Census Bureau, 2009).

The violent crime rate in 1985 was 2009/100,000, and this would decline marginally during the next two years before spiking in the early 1990s (FBI, Uniform Crime Reports, 2012). Property crime rates followed a similar trend, at 6565/100,000 in 1985, followed by a .8% decrease in 1986, and a 2.8% increase in 1987. Despite the apparent stability of the City's property crime problem during these three years, rates increased in each of the following eight consecutive years, peaking in 1995 at 10,300/100,000 (FBI, Uniform Crime Reports, 2012).

Sample

Heads of households were surveyed in 1987, and one year later in 1988, using a multi-stage, stratified, clustering sampling design (Shadish, Cook, & Campbell, 2002). Out of 237 ecologically defined and geographically stratified neighborhoods, 50 neighborhoods were randomly selected with a proportional probability to the population size for inclusion. Within these neighborhoods, street blocks were randomly sampled according to a probability proportional to street block population size (referenced via the number of residential, phone number listings for each street block) (Perkins, 1989). These narrowly-defined and small ecological units reduce the risk of systematic bias (Hipp, 2010). Fieldworkers sampled 12 households from each street block using an interval sampling procedure with a random start. Household heads, randomly sampled if necessary (i.e.-multiples heads), were designated respondents within households. The first wave of interviews was conducted in late winter and early spring of 1987 (N=412), while the follow-up interviews took place one year later (N=305) (Perkins, 1989).

The initial sampling frame was 601 respondents (Robinson, et al., 2003). Thirteen households were never used, and thirteen others were identified as abandoned/vacant. This leaves a total sampling frame of 575 households where interviews were attempted. If this number is used as the denominator, the response rate was 72% for Time1 (412/575). Analyses of on-site ratings of houses and streetblock characteristics showed no significant differences between houses where an interview was successfully completed and houses where an interview could not occur. One year later, 336 of the original 412 interviewees were available for re-interview. Seventy of the original 412 has moved off their streetblock, and were therefore ineligible for re-interview. Six others had passed on. Of these available 336, 305 were re-interviewed, for a response rate of 91% (Robinson, et al., 2003). Analyses were conducted to explore differences between the Time1 and Time2 samples. The samples showed to be very similar. At the Bonferroni adjusted alpha level, the two groups were not different on sex, age, race, or education. There was a small over-representation of homeowners (as compared to renters) in the Time2 sample. Accordingly, a weight was created and used in the analyses to account for this small bias.

Another issue common in survey research stems from non-generalizable samples due to complex sampling strategies and procedures (Tracy & Carkin, 2014). As such, we compared our sample with U.S. Census data from 1990 and found the groups to be similar. In the 1987 sample, the respective percentages of African Americans and Whites were 52% and 46%. Data from the 1990 census show 57% for African Americans and 41% for Whites. This pattern of decreasing White populations and therefore increasing proportions of African Americans accords with the changing demographics of Baltimore at the time

(Taylor, 2001). In additions, in the 1987 sample, 58% were homeowners and 42% were renters. By 1990, 52% were homeowners and 48% were renters. This is also a pattern we would expect given people that wealthier people, who were more likely to own, were moving out of the City, and the remaining concentration of poorer residents were more likely to rent.

At follow-up (Table 15), the sample was 45% White and 55% African American. Thirty-three percent of the respondents were male, and 64% had completed high school. The average age was 47 years; 58% of the sample were homeowners and the remaining were renters. Average household size was 2.9 persons. Approximately half the sample had a household income of less than $20,000 per year. The average length of residence in the neighborhood was 14.6 years, and the mean length of residence at the current address was 12.6 years (Perkins, 1989; Robinson et al., 2003). Compared to baseline, slightly more of the follow-up respondents were homeowners, and a weight was applied to control for this shift.

Dependent Variable

To measure the incivility variable, respondents' perceptions of incivilities on their respective street blocks were formed into an error-free latent index variable based on eight distinct items. With possible response categories of "big" (2)/ "somewhat" (1)/ and "not" (0), individuals estimated the degree to which each of the following neighborhood features presented a problem: (1) "Vandalism, like people breaking windows or spray painting buildings?"; (2) "Vacant housing?"; (3) "People who don't keep up their property or yards?"; (4) "People who say insulting things or bother other people when they walk down the street?"; (5) "Litter or trash in the streets?"; (6) "Vacant lots with trash or junk?"; (7) "Groups of teenagers hanging out on the street?"; and (8) "People fighting or arguing?" The use of an index variable to measure perceived incivilities follows previous work

Table 15
Summary Statistics-Individual and Street Block Variables

Variable	Mean	Std. Dev	Min	Max
Individual				
Race (1=white, 0=non-white)	0.46	0.50	0.00	1.00
Sex (1=male, 0=female)	0.33	0.47	0.00	1.00
Age	49.26	15.52	20.00	88.00
Marital Status (1=married, 0=other)	0.53	0.50	0.00	1.00
Work Full-Time (1=yes, 0=other)	0.49	0.50	0.00	1.00
Work Part-Time (1=yes, 0=other)	0.12	0.32	0.00	1.00
Children Living in Home (1=yes)	0.62	0.49	0.00	1.00
Occupational Prestige	38.80	23.93	6.00	93.00
Home Ownership (1=yes)	1.00	0.11	0.91	1.15
Baseline Risk	0.28	0.41	0.04	1.72
Follow-up Risk	0.22	0.37	0.03	1.66
Baseline Incivilities	0.40	0.35	0.06	1.58
Follow-up Incivilities	0.32	0.30	0.06	1.46
Street Block				
Baseline Risk	0.29	0.22	0.04	1.02
Baseline Incivilities	0.39	0.23	0.06	1.01
Robbery (logged)	6.55	1.00	3.9	8.42
Homeownership (proportion)	0.61	0.35	0	1.00

Notes: N= 305 residents of Baltimore City nested within 50 street blocks in 1987-1988. At the individual level, risk and incivilities were group-mean centered; age and occupational prestige were grand-mean centered. Risk and incivilities were grand-mean centered at the street block level. Robbery change variable was for the years 1985 through 1987 and was logged. Homeownership was weighted to account for fewer renters in follow-up sample.

by Rountree and Land (1996a and 1996b). Confirmatory structural equation modeling (SEM) was used and error-adjusted factor scores were computed to build the "error-free" latent variable. "Fighting" was the referent variable.

Independent Variables

Items from Ferraro's (1995) work were used to create crime risk perception as a latent index variable that contained four items gauging the perceived local probability of four crimes. Using response categories "big" (2)/ "somewhat" (1)/ "not" (0), individuals estimated how much of a problem (1) "Burglary?"; (2) "People selling illegal drugs?"; (3) "People getting robbed on the street?"; and (4) "People getting assaulted or beaten up on the street?" were on their particular street block. We found that using this constructed variable, rather than a dichotomous variable indicating whether an individual feels safe or not on her street block (Rountree & Land, 1996a; 1996b), provides a more diversified measure of perceived risk, allowing for more variability in this variable, and producing more efficient estimates of the effect of perceived risk on the change in future incivilities. Further, another benefit of our variable is that it is measured at the street block level—a relatively small unit of measurement that is not as susceptible to the threats to validity present when measuring at higher levels of aggregation such as the census tract (Hipp, 2010). In terms of our proposed theory, an individual seeing more local crime risk than his or her neighbors may be more tuned in to, or more likely to give weight to, later shifts in incivilities on his or her street block. For this reason, we group-mean centered perceived risk to gauge differences between people on the same street block (see Firebaugh, 1980 for more on group-mean centering).

Additionally, BWT argues that elements at street block and neighborhood levels have impacts on individual perceptions. As such, we also controlled for the baseline perceived risk at the street block level by using street block mean scores based on the average of the individual-level risk index scores (before group-mean centering) to capture street block variations in perceived risk. These scores were entered grand-mean centered to compare differences across street blocks.

Covariates

Perceptions of incivilities at the individual level and at baseline was entered as a control variable. The benefit of controlling for the outcome at baseline is that the remaining variation of incivilities at follow-up reflects unexpected changes in incivilities perceptions; in other words, the inclusion of this variable in the model allows for a lagged analysis of earlier factors on later changes in incivilities. The eight-item index was constructed identically to the error-free latent index outcome variable. Because of the possibility of omitted variable bias for the estimated effect of perceived risk on perceived incivilities from not including a control for crime, the logged neighborhood robbery rates per 100,000 in Baltimore from the two years prior to the beginning of the surveys (1985-1986) and for the first year of the survey (1987) were included in the final model.

Demographics

Harcourt (2001) argued that perceptions of disorder are variable and are influenced by many individual factors. As such, and to limit the amount of bias in the estimate(s) of interest, it was necessary to control for individual demographic variables. Indeed, recent research shows individual factors have significant effects on the perception of disorder (Hipp, 2010; Wallace, 2011). Perception of disorder has shown to significantly decrease with age (Wallace, 2011), but not just because, all else equal, an older individual will perceive less incivility than a younger person. Rather, the lifestyles of an older individual versus a younger individual are different, leading to differences in how disorder is perceived. Therefore, age was included as a lifestyle proxy and grand-mean centered. Along the same lines, occupational prestige was entered as an indicator of socioeconomic status (Duncan, 1961), and was grand-mean centered. A number of dichotomous demographic variables were also included in our models. A gender indicator was included to account for the safety concerns that may cause females to be more aware of dangers in the environment (Hipp, 2010). This variable was coded "0" (females) and "1" (males). Background effects were controlled for with a race variable that was coded "0" for non-white and "1" for white. A neighborhood attachment variable (homeowner) was included to control for community attachment. This variable was coded "0" (renter household) and "1" (owner occupied). To account for altruistic increase in incivilities associated with being responsible for the safety of dependent children a variable that captured if children resided in the home "1" or not "0" was included in the model. Further, a married variable was also included and coded "1" for married individuals and "0" for all others. Last, two employment dummy variables were included for full-time and part-time workers. These variables captured working full-time "1" or not "0" and working part-time "1" or not "0", respectively.

Analytic Strategy

Hierarchical linear models (Hox, 2010; Raudenbush & Bryk, 2002) permitted separating psychological from ecological impacts or risk and incivilities perceptions (Taylor, 2010). Four models are presented here: Model A, an analysis of variance (ANOVA) in HLM, assessed whether significant ecological variation in the outcome incivilities perceptions appeared. Model B (ANCOVA) entered baseline perceived incivilities to learn whether significant ecological variation remained in changes in incivilities between baseline and follow-up. Model C (ANCOVA) added risk perceptions at both the individual (L1) and street block (L2) levels (pathways *e* and *b*, respectively) permitting the estimation of each on changes in perceived incivilities.

Model D was a random coefficients regression (RCR) HLM model that controlled for compositional effects (pathway *h*) and recent neighborhood robbery rates (pathway *d*). While finding the average effect of the individual's perception of risk on later incivilities is important in an important test of this reversed model, understanding how ecological and psychological variables might moderate that effect gives a much more comprehensive understanding of risk perceptions' impact on incivilities, and gives us clues about causal mechanisms. Therefore, in Model E, we estimated an intercepts-and-slopes-as-outcomes (IASAO) model that allowed the effect of individual risk perceptions on incivilities to vary and tested whether street block stability significantly moderates risk and incivilities perceptions. Model F was an RCR model that tested the traditional BWT pathway.

In Model G, following Hipp (2010) and Wallace (2011), we used a fixed effect model to control for unobservable differences on each street block that may bias the estimation of the resulting coefficients. In the fixed effect model, a dummy variable representing every street block but one is included in the model. This effectively absorbs all of the variation between each street block. This results in coefficient estimates that are found using only the variation of individuals within the same street block, producing consistent results albeit with the sacrifice of less efficiency. Thus, in this model, any differences in incivility perceptions among the respondents are due to individual differences and are not a function of the street block environment (Halaby, 2004; Wallace 2011).

Results

Model A (table not shown) revealed significant ecological variation in perceived incivilities and that over a third of the variation in incivilities perceptions is ecological (x^2 = 207.596, $p < .001$, *ricc* = .350).

As expected, Model B (Table 16) revealed a significant impact of baseline on follow-up incivilities perceptions. Significant ($p < .01$) ecological variation remained in the outcome, which is now changes in perceived incivilities after the baseline interview.

In Model C (ANCOVA), results showed that both individual and street block differences in baseline risk perceptions shaped later changes in incivilities perceptions ($p < .001$ and $p < .05$, respectively). Those respondents perceiving more risk than their neighbors were more likely later to see local deterioration emerging. In addition, street blocks where residents initially saw elevated risk were also more likely to see more deteriorated conditions emerging subsequently ($p < .001$). Even after factoring in the multilevel impacts of perceived risk on later changes in incivilities perceptions, significant ($p < .05$) ecological variation in the latter persisted.

Model D (RCR) controlled for demographics/compositional effects at the individual level, and a local crime rate variable (robbery) at the street block level. Both the ecological and psychological impacts of perceived risk remained significant. Standard errors of coefficients were clustered around street block to account for autocorrelations of the independent variables of neighbors living on the same street block.

Model E (IASAO) results showed that the impact of risk perceptions indeed varied significantly across street blocks (χ^2 = 65.22, p = .014) and could be predicted by homeownership. An indicator of stability, the proportion of street block homeownership (0 – 1) was entered into the model individually for each observation and was able to significantly predict the varying slope of risk on incivilities perceptions. The intercept for risk perceptions at level one was .197 and the impact of a resident's street block homeownership proportion on the slope of risk predicting incivilities perceptions was -.202. This estimation led to an interesting theoretical interpretation: as a street block's score on stability increases, impact of perceived risk on perceived incivilities decreases. We conclude that residents who live on a block with higher stability will metaphorically wear rose-colored glasses and be less likely to associate heightened perceptions of risk to heightened perceptions of unexplained incivilities. In this model ecological variation on the outcome was now no more than sampling error.

Model F (RCR) (Table 17) shows the results for testing a segment of the traditional BWT model (incivilities→risk). Here we find partial support. While no evidence is found for perceived incivilities shaping later perceived crime risk at the individual level, we do find an ecological effect—incivilities at the street block level predict later changes in risk perceptions.

In the fixed effect analyses (Table 18), we accounted for possible bias in the coefficient estimates due to persistent, unexplained neighborhood variables that account for correlation in neighbors' responses in the explanatory and explained variables. We find that perceived risk still significantly predicts future unexplained incivilities, though the estimated coefficient on Time1 risk drops from .197 in Model E to .118 in Model G (see Table 18). This finding highlights that perceived risk's effect on future unexplained incivilities is not due to similarities in variables from living on the same street block as your neighbors. The neighborhood robbery rate was dropped from the model due to its lack of variability at the street block over time. Now we turn to the results where we tested part of BWT by using the fixed effect modeling to see if perceived incivilities had a significant effect on future unexplained risk perceptions after controlling for all other factors including 49 dummy variables for each street block. We find no significant effect (see Table 18).

Table 16
ANCOVA, RCR, and IASAO Models of Incivilities Perceptions

Variables	Model B	Model C	Model D	Model E
Individual Level				
Time 1 Incivilities	0.618 ***	0.501***	0.507 ***	0.519 ***
S.E. T1 Incivilities	(0.047)	(0.066)	(0.064)	(0.063)
Time 1 Risk		0.094*	0.090 *	0.197*
S.E. T1 Risk		(0.049)	(0.048)	(0.084)
% Homeownership				-0.202*[i]
S.E. of % Home.				(0.113)
Marital Status			-0.014	-0.007
Homeownership			-0.039	-0.051*
Gender			-0.033	-0.028
Work Full-Time			0.003	0.005
Work Part-Time			-0.035	-0.050
Race			-0.001	-0.001
Age			-0.001	-0.001
Children Living at			-0.018	-0.005
Occupational Prestige			-0.001	-0.001
Streetblock Level				
Time 1 Risk		0.311***	0.279 **	0.269**
S.E. T1 Risk		(0.080)	(0.077)	(0.077)
Robbery			-0.004	-0.004

*[i] $p < .05$, one-tailed; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Notes*:* N= 305 residents of Baltimore City nested within 50 street blocks in 1987-1988. At the individual level, risk was group-mean centered; age and occupational prestige were grand-mean centered. Risk was grand-mean centered at the street block level. Robbery change variable was for the years 1985 through 1987 and was logged. Homeownership was weighted to account for fewer renters in follow-up sample.

Table 17
HLM Model of Traditional BWT model

Variables	Model F
Individual level	
Time 1 Incivilities	-0.005
S.E. of T1 Incivilities	0.077
Time 1 Risk	0.603***
S.E. of T1 Risk	0.084
Marital Status	0.044
Homeownership	-0.054
Gender	0.010
Work Full-Time	0.033
Work Part-Time	-0.025
Race	0.071
Age	0.000
Children Living at Home	-0.044
Occupational Prestige	0.000
Streetblock level variables	
Time 1 Incivilities	0.233*
S.E. of T1 Incivilities	0.103
Robbery	-0.012

*p < .05; **p < 0.01; ***p < 0.001

Notes: N= 305 residents of Baltimore City nested within 50 street blocks in 1987-1988.At the individual level, incivilities was group-mean centered; age and occupational prestige were grand-mean centered. Incivilities was grand-mean centered at the street block level. Robbery change variable was for the years 1985 through 1987 and was logged. Homeownership was weighted to account for fewer renters in follow-up sample.

In summary, intra- and inter-street block differences in perceived risk at baseline significantly shaped later changes in perceived incivilities, even after controlling for compositional differences and neighborhood crime. Street block residents' perceptions of later shifts in local deteriorated conditions were significantly shaped by the entire street block's perceptions of crime risk, and by residents' perceptions of crime risk relative to their on-block neighbors. This impact varied significantly by street block. Those blocks that had a high percentage of homeownership felt virtually no impact of crime risk on later neighborhood problems, while street blocks that had low homeownership felt double the average impact of risk. Testing the traditional BWT model, we find support for the ecological component but not for the individual-level, psychological model. Finally, accounting for possible bias in the coefficient estimates due to omitted street block-level variables, the fixed effects model confirms the individual-level results of our HLM models.

Discussion

Robinson et al. (2003) speculated that the causal ordering suggested by the incivilities thesis might have been viewed too narrowly: "It also seems plausible the connection may be working the other way, with changing fear and changing satisfaction driving changes in perceived problems". Perkins and Taylor (1996) and Taylor (1999) have mentioned that incivilities, risk, and fear may be ordered and related in ways we do not yet understand fully. Building on these ideas, this work targeted a key version of the incivilities thesis—the "broken windows" version—and tested a central longitudinal link in that version—the incivilities-risk link. But it also tested a model assuming a temporal ordering opposite that suggested by Wilson and Kelling (1982). Results more strongly supported this reversed ordering: earlier risk, operating through two pathways, altered

Table 18

Fixed Effects Model of Incivilities Perceptions

	Model G	
	Dependent Variable	
Independent Variables	Time 2 Incivilities	Time 2 Risk
Time 1 Incivilities	0.442***	-0.0118
S.E. of T1 Incivilities	0.0690	0.0869
Time 1 Risk	0.118*	0.599***
S.E. of T1 Risk	0.0520	0.0819
Marital Status (1=married 0=other)	0.0206	0.0833
Homeownership (1= yes 0=no)	-0.0198	-0.0512
Gender	-0.0398	0.0113
Work Full-Time	0.0195	0.0160
Work Part-Time	-0.0165	-0.0362
Race	0.0277	0.1065
Age	-0.0003	-0.0004
Children Living at Home	0.0031	-0.0139
Occupational Prestige	0.0001	0.0002
F-test (d.f. 60, 242)	15.52	6.94
$p<$	0.001	0.001
R^2	0.6805	0.6336

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Notes: N= 305 residents of Baltimore City nested within 50 street blocks in 1987-1988. Analysis does not show coefficients for 49 dummies representing all but one streetblock. Robbery change variable was for the years 1985 through 1987 and was logged. Homeownership was weighted to account for fewer renters in follow-up sample later incivilities perceptions. This is the first empirical test, of which we are aware, finding multilevel, lagged impacts of risk perceptions on later shifts in local perceived incivilities.

Wilson and Kelling's influential 1982 piece argued that incivilities in communities are identified and interpreted by residents and passersby as an indication of imminent breakdowns in resident-based informal control over people and events in their immediate surrounds. This results in increased perceptions of threat to personal safety—ultimately causing law-abiding people to withdraw from common residential spaces. Underlying this contention is the assumption that residents in fact see and interpret cues—specifically incivilities—similarly. Wilson and Kelling quote Glazer (1979) to support their contention: "…the proliferation of graffiti, even when not obscene, confronts the subway rider with the inescapable knowledge that the environment he must endure for an hour or more a day is uncontrolled and uncontrollable, and that anyone can invade it to do whatever damage and mischief the mind suggests". Our results do not support such a view.

Rather, results support an alternative view that crime risk perceptions themselves may shape how one defines features within his or her environment. Individuals who feel more at risk earlier are likely to define features in his or her immediate surroundings as more problematic later. Further, an ecological dynamic also was operative: those on street blocks where average perceived risks were higher initially reported more local deterioration one year later. These findings potentially suggest two pathways for how individual incivilities perceptions might be framed. First, whereby an individual-level, psychological process produces these perceptions based on feelings of crime risk, and second, a pathway whereby something in the street block ecology, possibly discussion or interaction with other street block residents, shape how ambiguous cues are defined. We argue that these empirical findings fit within a symbolic interactionist framework that emphasizes how individuals perceive and define interactions and experiences in ways that are non-uniform. Thus, our findings are consistent with Harcourt's (2001) hypothesis that defining disorder is a subjective process with considerable variation.

The above findings advance recent research in disorder theory that questions the existence of universally accepted cues of physical and social incivilities. Whether one decides to define a cue as disorderly is dependent upon many factors and "biases", including: race, place, social cohesion, neighborhood attachment, engagement in protective behaviors, and routine activities (Carvalho & Lewis, 2003; Franzini et al., 2008; Hipp, 2010; Jackson, 2004; Sampson & Raudenbush, 2004; Wallace, 2011; Wickes et al., 2013). Now we have evidence that individual perceptions of crime risk significantly shape perceptions of local conditions as well. People who feel a high, localized risk of crime are "biased" toward seeing possibly ambiguous features of a neighborhood as problematic. Thus, disorder theory has evolved to the point that fundamentally challenges the beginning link of the incivilities thesis—that people see and define their surroundings similarly.

Our results also contribute to the literature on neighborhood stability and reactions to crime and related phenomena (Taylor, 1996). Out of many possible demographic, crime, and neighborhood structure variables, we found that only street blocks with higher proportions of homeownership were less likely to react to crime risk by seeing more neighborhood problems later. In this way, homeownership—as an indicator of residential stability—appeared to insulate certain street blocks from the untoward consequences of feeling crime risk. It seems plausible that residents of high homeownership areas would tend to define their surroundings as better than those who are in high instability neighborhoods. This finding supports the stability emphasis of Bursik and Grasmick's (1993) systemic model of informal social control. They write, "At the private level, ongoing changes in the residential population of a neighborhood make it very difficult to establish and maintain intimate primary ties within the community". It is within these private levels (Baltimore street blocks) where ongoing changes in homeownership were high that individuals were more likely to see their blocks deteriorating. In addition, these findings support Taylor's (1996) work that showed stability to have expected positive impacts on involvement, attachment, and responses to disorder. He captures the importance of stability succinctly: "More stable neighborhoods have residents who care more about their community. But in addition…the work here suggests that because stability enhances neighbors' knowing one another and caring about their neighborhood, it also makes them feel less vulnerable to crime and related problems". In sum, homeownership showed to be the only item that could predict why risk perceptions had a stronger impact on later neighborhood deterioration in some areas but not others. To our knowledge, this work is the first to identify this empirical phenomenon.

Conclusions

Using panel data, we found empirical support for both Broken Windows (Wilson & Kelling, 1982) and a reversed Broken Windows theory. Although, in line with recent research on how disorder and incivility perceptions are indeed socially constructed, our results more strongly favored the reversed Broken Windows model. Two different types of models found strong evidence that crime risk perceptions have lagged impacts on changes in perceived neighborhood conditions. Further, the impact of the effect is contingent on the degree of stability on a particular street block—blocks with low stability saw high impacts of risk and blocks with high stability were virtually inoculated against any potential impact.

NOTES

[1]Also referred to as multi-year teaching/placement, continuous learning, or family-style learning. The term “Looping” will be used primarily throughout this paper.

[2]Before 2006, the person identified as the student’s classroom teacher is the person that administered the exam. While this is very likely to be the student’s actual classroom teacher, in many cases it was not. In some schools tests were administered in large sections of multiple classrooms or students switched to another classroom during testing.

[3]This includes codes for disabilities that greatly affect the learning process such as autism, deafness, developmental delay, serious emotional disability, intellectually disabled, orthopedic impairment, speech impairment, traumatic brain injury, and visual impairment.

[4] Self-contained classrooms are such that students have the same classroom teacher for the core curriculum subjects. These subjects include but are not limited to history, science, writing and the subjects of interest for this research, math and reading. Self-Contained classes do not include special education classes. A useful restriction because the curricular and student-teacher interactions in these classrooms significantly differs from that of a traditional classroom settings.

[5]About half of these observations are eliminated because when a new student entered a class, an updated observation was created for *every incumbent student in the class*

[6]While excluding gifted students from the regression analysis will help with any non- random sample selection associated with this variable, it does not control for any peer effects other students may experience from having gifted students in their class.

Rather, peer effects are controlled for by including lagged average class scores in math and reading

[7]Students in a classroom of a teacher they have had previously, include all students that are repeating with a teacher for a second year as well as students that are repeating with a teacher for a third and fourth year

[8]Controlling for peer-effects is necessary when students are not randomly assigned to classroom and peer-group composition effects achievement. Clotfelter, Ladd, and Vigdor (2006) use an alternative approach to addressing this problem by studying students that are randomly assigned to classrooms.

[9]The summation series in equation (2) is over grade level or years of schooling

[10]The parameters that specify each grade level achievement function are not specific to each student, so the marginal effects are the same for all students in a given grade at the same time.

[11]See Chetty et al. (2014) for the most convincing argument as to the lack of bias in VAMs' estimates when $0 < \lambda < 1$. Their models actually use a cubic-polynomial of prior years' test scores in both math and reading. This research extends the baseline model to replicate this approach.

[12]If this assumption does not hold and student heterogeneity is correlated with a student's probability of repeating with a teacher than student fixed effects are necessary to minimize bias in the estimation.

[13]This includes a square of age to allow for diminishing returns of age on achievement

[14]This includes indicator variables for American Indian, Hispanic, Black, Multi-racial, Pacific Islander, and White ethnicities.

[15]I again allow for diminishing returns of attendance on achievement.

[16]This includes separate variables for when the student moved within the same school district and when the student moved to another school district.

[17]For maximum flexibility, lags include cubics of the class averages in math and reading test scores.

[18]Within the same school, grade, and year, teachers with estimated effects on achieve- ment that are one standard deviation (or more) above the mean teacher (normalized to 0) are considered to be "high" quality

[19]see E. A. Hanushek and Rivkin (2010) for a summary.

[20]As with math score teacher estimates, I restrict estimation of teacher effects to teachers with more than 20 student observations.

[21]See (B. A. Jacob & Lefgren, 2007) for information on what teacher attributes parents prefer.

[22]Results not reported. See author for results

[23]If students are randomly sorted to classrooms, then the true effect of next year's teacher on a student's achievement this year should be equal to zero

[24]sometimes referred to as soft skills, personality traits, non-cognitive traits, non-cognitive abilities, character, and socio-emotional skills

[25]see Kautz et al. (2014), Heckman and Kautz (2012), Borghans, Duckworth, Heckman, and Ter Weel (2008)

[26]See Chapter 1 and the analogous derivation using a student's lagged achievement score.

[27]Self-contained classrooms are such that students have the same classroom teacher for the core curriculum subjects and do not include special education classrooms.

[28]Even if every teacher a student is taught by in a school year is properly identified (not always the case), to find the true teacher effects would require controlling for any externalities of all other contemporaneous teachers.

[29]Students in a classroom of a teacher they've had previously, include all students that are repeating with a teacher for a second year as well as students that are repeating with a teacher for a third and fourth year

[30]These include student's age; ethnicity; gender; exceptionality codes for disabilities that greatly affect the learning process such as Autism, deafness, developmental delay, serious emotional disability, intellectually disabled, orthopedic impairment, speech impairment, traumatic brain injury, and visual impairment; socio-economic disadvantage; days absent; limited English proficiency; and academic giftedness

[31]Student subjective assessment of achievement is based on anticipated grades which is determined by how well a student perceives she has done throughout the year while teacher subjective assessments are based on how well the teacher thinks the student has done. Both variables are specified before the student has received her state-administered test outcome.

[32]See Chapter 1's Data section for a more extensive explanation of the use of third grade pre-tests.

[33]Technically there is a very limited number of third graders included in the sample. For these students information is available from the previous year because they are repeating the third grade included in the sample used for the reported results

[34]This is a fundamental assumption when using Value-Added Models to estimate teacher effects, however, it has been the subject of great debate. See (Rothstein, 2017) and (Chetty, Friedman, & Rockoff, 2016) for the most up to date arguments.

[35]see Chapter 1

[36]see Chapter 1

[37]Conditional on school, grade, and year, teachers that have a value-added estimate $\geq$ one standard deviation of the distribution of estimated teacher effects on achievement are considered high-quality teachers

[38]The values of the outcomes on time spent on homework are on a *weekly* basis. They are measured as follows: O-no time spent on homework, 1-less than an hour spent on homework, 2-1 to 3 hours, 3-3 to 5 hours, 4-5 to 10 hours, and 5-10 or more hours.

[39]Not reported. Contact author for results

[40]Significance at 99% level is maintained.

BIBLIOGRAPHY

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of labor Economics*, *25*(1), 95–135.

Alloway, T. P., Doherty-Sneddon, G., & Forbes, L. (2012). Teachers' perceptions of classroom behaviour and working memory. *Educational Research and Reviews*, *7*(6), 138.

Altonji, J. G., Conley, T., Elder, T. E., & Taber, C. R. (2010). Methods for using selection on observed variables to address selection on unobserved variables.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005a). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human resources*, *40*(4), 791–821.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005b). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political Economy*, *113*(1), 151–184.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2008). Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization. *The American Economic Review*, *98*(2), 345–350.

Arcidiacono, P., Foster, G., Goodpaster, N., & Kinsler, J. (2012). Estimating spillovers using panel data, with an application to the classroom. *Quantitative Economics*, *3*(3), 421–470.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in los angeles* (Tech. Rep.). National Bureau of Economic Research.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics*, *29*(1), 37–65.

Baltagi, B. (2008). *Econometric analysis of panel data*. John Wiley & Sons.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, *70*(5, Part 2), 9–49.

Becker, G. S. (1994). Human capital revisited. In *Human capital: A theoretical and empirical analysis with special reference to education (3rd edition)* (pp. 15–28). The University of Chicago Press.

Bedard, K., & Ferrall, C. (2003). Wage and test score dispersion: some international evidence. *Economics of education review*, *22*(1), 31–43.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in mexican high schools. *Journal of Political Economy*, *123*(2), 325–364.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of political economy*, *75*(4, Part 1), 352–365.

Ben-Porath, Y. (1970). The production of human capital over time. In *Education, income, and human capital* (pp. 129–154). NBER.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*(4), 327 - 336. Retrieved from http://www.sciencedirect.com/science/article/pii/S1041608011000124

Blazar, D. (2015). Grade assignments and the teacher pipeline: A low-cost lever to improve student achievement? *Educational Researcher*, *44*(4), 213–227.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of education*, 113–121.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of human Resources*, *43*(4), 972–1059.

Boudett, K. P., & Murnane, R. J. (2013). Data wise: A step by step guide to using assessment results to improve teaching and learning. *Harvard Education Press*.

Brown, D. F. (2004). Urban teachers' professed classroom management strategies reflections of culturally responsive teaching. *Urban Education*, *39*(3), 266–289.

Brunton-Smith, I. (2011). Untangling the relationship between fear of crime and perceptions of disorder. *British Journal of Criminology* 51, 885-99.

Brunton-Smith, I. & Sturgis, P. (2011). Do neighborhoods generate fear of crime? An empirical test using the British Crime Survey. *Criminology* 41, 331-69.

Bull, R., & Scerif, G. (2001, Jun). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293. (dummy note)

Caauwe, C. M. (2009). The impact of looping practices on student achievement at a Minnesota inner city elementary school: A comparison study. *Online Submission*.

Caeyers, B., & Fafchamps, M. (2016). *Exclusion bias in the estimation of peer effects* (Tech. Rep.). National Bureau of Economic Research

.

Campbell, R., Kyriakides, L., Muijs, R., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, *29*(3), 347–362.

Card, D. (1999). *The casual effect of education on earnings w o. ashenfelter, d. card handbook of labour economics.* North Holland. Amsterdam.

Carvalho, I. & Lewis, D.A. (2003). Beyond community: Reactions to crime and disorder among inner-city residents. Criminology 41, 779-812.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *The American Economic Review*, *95*(4), 1237–1258.

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Discussion of the american statistical asso- ciation's statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy*, *1*(1), 111–113.

Chetty, R., Friedman, J., & Rockoff, J. (2017). Measuring the impacts of teachers: Reply to rothstein1.

Chetty, R., Friedman, J. N., & Rockoff, J. (2016). Using lagged outcomes to evaluate bias in value-added models. *The American Economic Review*, *106*(5), 393–399.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *The American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014c). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633–2679.

Chetty, R., & Hendren, N. (2015). The impacts of neighborhoods on intergenerational mobility: Childhood exposure effects and county-level estimates.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *The American Economic Review*, *106*(4), 855–902.

Chudgar, A., & Quin, E. (2012). Relationship between private schooling and achievement: Results from rural and urban India. *Economics of Education Review*, *31*(4), 376–390.

Cistone, P., & Shneyderman, A. (2004). Looping: An empirical evaluation. *International Journal of Educational Policy, Research, and Practice: Reconceptualizing Childhood Studies*, *5*(1), 47–61.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources*, *41*(4), 778–820.

Collins, S. (2010). Multiyear student/teacher relationships and language development in children of Hawaiian descent at Kamehameha schools community based early child- hood education program.

Cook, F. L. and Skogan, W.G. (1984). Evaluating the changing definition of a policy issue in Congress: Crime against the elderly. In H. R. Rodgers (Ed.) Public Policy and Social Institutions. Greenwich: JAI Press.

Cornelissen, T., et al. (2008). The stata command felsdvreg to fit a linear model with two high-dimensional fixed effects. *Stata Journal*, *8*(2), 170.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883–931.

Currie, J., & Thomas, D. (2001). Early test scores, school quality and ses: Longrun effects on wage and employment outcomes. In *Worker wellbeing in a changing labor market* (pp. 103–132). Emerald Group Publishing Limited.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics*, *5*(2), 29–57.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American economic review*, *95*(2), 158–165.

De Witte, K., & Kortelainen, M. (2013). What explains the performance of students in a heterogeneous environment? conditional efficiency estimation with continuous and discrete environmental variables. *Applied Economics*, *45*(17), 2401–2412.

De Witte, K., & Van Klaveren, C. (2014). How are teachers teaching? a nonparametric approach. *Education Economics*, *22*(1), 3–23.

Dieterle, S., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do principals assign students to teachers? finding evidence in administrative data and the implications for value added. *Journal of Policy Analysis and Management*, *34*(1), 32–58.

Dugan, D. J., Jamison, D. T., & Radner, R. (1976). Scholastic achievement: its determinants and effects in the education industry. In *Education as an industry* (pp. 53–94). NBER.

Duncan, O. D. (1961). A socioeconomic index for all occupations. In A. J. Reiss (Ed.) Occupations and Social Status. New York: Free Press.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44–52.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school-and teacher-level models in missouri. *Statistics and Public Policy*, *1*(1), 19–27.

Ehrenberg, R. G., & Brewer, D. J. (1994). Do school and teacher characteristics matter? evidence from high school and beyond. *Economics of Education Review*, *13*(1), 1– 17.

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? evidence from the national educational longitudinal study of 1988. *Industrial & Labor Relations Review*, *48*(3), 547–561.

Elliott, J. G. (2009). The nature of teacher authority and teacher expertise. *Support for Learning*, *24*(4), 197–203.

Federal Bureau of Investigation. (2012). *Uniform Crime Reports*. Prepared by the National Archive of Criminal Justice Data. Retrieved April 12, 2012.http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012.

Feng, L., & Sass, T. R. (2016). Teacher quality and teacher mobility. *Education Finance and Policy*.

Ferraro, K. F. (1995). *Fear of crime: Interpreting victimization risk*. SUNY Press.

Finder, A. (2005). Goodbye, class. see you in the fall. *New York Times*, *154*, 2.

Firebaugh, G. (1980). Groups as contexts and frog ponds. In K. Roberts & L. Burstein (Eds.) *Issues in Aggregation*. San Francisco: Jossey-Bass.

Fletcher, J. M., Horwitz, L. I., & Bradley, E. (2014). *Estimating the value added of attending physicians on patient outcomes* (Tech. Rep.). National Bureau of Economic Research.

Franzini, L., Caughy, M. O. Nettles, S. M., & O'Campo, P. (2008). Perceptions of disorder: Contributions of neighborhood characteristics to subjective perceptions of disorder. *Journal of Environmental Psychology,* 28, 83-93.

Fryer Jr, R. G. (2016). *The'pupil'factory: Specialization and the production of human capital in schools* (Tech. Rep.). National Bureau of Economic Research.

Gallagher, N. A., Gretebeck, K. A.., Robinson, Torres, E. R., Murphy, S. L., & Martyn, K. K. (2010). Neighborhood factors relevant for walking in older, urban, African American adults. *Journal of Aging and Physical Activity*, 18(1), 99-115.

Garofalo, J., & Laub, J. (1978). The fear of crime: Broadening our perspective. *Victimology*, 3, 242-53.

Gau, J. M., & Pratt, T. C. (2008). Broken windows or window dressing? Citizens' (in)ability to tell the difference between disorder and crime. *Criminology and Public Policy,* 7, 163-94.

Gaustad, J. (1998). Implementing looping. eric digest, number 123.

Glazer, N. (1979). On subway graffiti in New York. *The Public Interest,* 54, 3-12.

Goldhaber, D., & Chaplin, D. (2015). Assessing the "rothstein falsification test": Does it really show teacher value-added models are biased? *Journal of Research on Educa- tional Effectiveness*, *8*(1), 8–34.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *The American Economic Review*, 250–255.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, *1*(1), 28–39.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 505–523.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? high school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying effective teachers using performance on the job. the hamilton project policy brief no. 2006-01. *Brookings Institution*.

Grant, J., et al. (1996). *The looping handbook: Teachers and students progressing together.* ERIC.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, *40*(2), 190–222.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*.

Hampton, F. M., et al. (1997). Enhancing urban student achievement through multi-year assignment and family-oriented school practices. *ERS spectrum*, *15*(2), 7–15.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology,* 30, 507-44.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, *61*(2), 280–288.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of economic literature*, *24*(3), 1141–1177.

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.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of applied econometrics*, *18*(5), 527–544.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (Tech. Rep.). National Bureau of Economic Research.

Hanushek, E. A., Kain, J. F., Rivkin, S. G., & Branch, G. F. (2007). Charter school quality and parental decision making with school choice. *Journal of public economics*, *91*(5), 823–848.

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American economic review*, 1184–1208.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051–1078.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, *4*(1), 131–157.

Harcourt, B. E. (2001). *Illusion of order: The false promises of Broken Windows Policing*. Cambridge: Harvard University Press.

Harcourt, B. E., & Ludwig, J. (2006). Broken windows: New evidence from New York City and a five-city social experiment. *University of Chicago Law Review,* 1, 271-320.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*(1), 73–112.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, *95*(7), 798–812.

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*, 183–204.

Heckman, J. J., Humphries, J. E., & Kautz, T. (2014). *The myth of achievement tests: The GED and the role of character in American life*. University of Chicago Press.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, *19*(4), 451–464.

Hegde, A. V., & Cassidy, D. J. (2004). Teacher and parent perspectives on looping. *Early Childhood Education Journal*, *32*(2), 133–138.

Hitz, M. M., Somers, M. C., & Jenlink, C. L. (2007). The looping classroom: Benefits for children, families, and teachers. *YC Young Children*, *62*(2), 80.

Hipp, J. (2010). Resident perceptions of crime and disorder: How much is “bias,” and how much is social environment differences? *Criminology*, 48, 475-508.

Hox, J. J. (2010). *Multilevel Analysis (2nd ed.).* East Sussex: Routledge.

Hoxby, C. M., Rockoff, J. E., of Economics, H. U. D., & of Business, C. U. G. S. (2004). *The impact of charter schools on student achievement*. Dept. of Economics, Harvard University.

Hunter, A. (1978). Symbols of incivility: Social disorder and fear of crime in urban neighborhoods. Paper presented at the 1978 meeting of the American Society of Criminology, Dallas, TX, November 8-12.

Innes, M. (2004). Signal crimes and signal disorders: Notes on deviance as communicative action. British Journal of Sociology, 55, 337-55.

Innes, M. (2005). What’s your problem? Signal crimes and citizen-focused problem solving. *Criminology and Public Policy*, 4, 187-200.

Irving, M. A. (2008). In the loop: An examination of the effectiveness of looping for african american students.

Jackson, J. (2004). Experience and expression: Social and cultural significance in the fear of crime. *British Journal of Criminology,* 66(6), 946-966.

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics*, *27*(2), 213–256.

Jackson, C. K. (2013a). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*.

Jackson, C. K. (2013b). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. *NBER Working Paper 18624*.

Jackson, C. K. (2016). *What do test scores miss? the importance of teacher effects on non-test score outcomes* (Tech. Rep.). National Bureau of Economic Research.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4).

Jackson, C. K., Rockoff, J. E., Staiger, D. O., et al. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, *6*(1), 801–825.

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *The Journal of Economic Perspectives*, *30*(3), 85–107.

Jacob, B. A., & Lefgren, L. (2007). What do parents value in education? an empirical investigation of parents' revealed preferences for teachers. *The Quarterly Journal of Economics*, *122*(4), 1603–1637.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics*, *26*(1), 101–136.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human resources*, *45*(4), 915–943.

Jacob, B. A., & Rockoff, J. E. (2012). Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments. *The Education Digest*, *77*(8), 28.

Jing-Qiu, L. (1997). The emotional bond between teachers and students: Multi-year rela- tionships. *Phi Delta Kappan*, *79*(2), 156.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. research paper. met project. *Bill & Melinda Gates Foundation*.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? evidence from New York City. *Economics of Education review*, *27*(6), 615–631.

Kane, T. J., & Staiger, D. O. (2001). *Improving school accountability measures* (Tech. Rep.). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Tech. Rep.). National Bureau of Economic Research.

Kane, T. J., Staiger, D. O., Grissmer, D., & Ladd, H. F. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings papers on education policy*(5), 235–283.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success* (Tech. Rep.). National Bureau of Economic Research.

Kelley, C. (2004). The impact of looping on reading achievement. *Retrieved May, 15, 2011*

Kelly, G. A. (1963). *A Theory of Personality: The Psychology of Personal Constructs.* New York: W. W. Norton.

Kenny, L. W. (1982). Economies of scale in schooling. *Economics of Education Review*, *2*(1), 1–24.

Kerssen-Griep, J., Gayle, B. M., & Preiss, R. W. (2006). Classroom interaction and educational outcomes. *Classroom communication and instructional processes: Advances through meta-analysis*, 129–147.

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, *3*(2), 333–362.

Kinsler, J. (2016). Teacher complementarities in test score production: Evidence from primary school. *Journal of Labor Economics*, *34*(1), 29–61.

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*(3), 560–572.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education review*, *28*(6), 682–692.

Koedel, C., & Betts, J. (2010). Value added to what? how a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, *5*(1), 54–81.

Koedel, C., & Betts, J. R. (2007). Re-examining the role of teacher quality in the educational production function. *University of Missouri Department of Economics Work- ing Paper*, 07–08.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? an extended analysis of the rothstein critique. *Education Finance and Policy*, *6*(1), 18–42.

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models. *B.E. Journal of Economic Analysis and Policy*, *12*(1).

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Koedel, C., & Tyhurst, E. (2012). Math skills and labor-market outcomes: Evidence from a resume-based field experiment. *Economics of Education Review*, *31*(1), 131–140.

Krueger, A. B. (2002). Understanding the magnitude and effect of class size on student achievement. *The class size debate*, 7–35.

Kubrin, C. E. (2008). Making order of disorder: A call for conceptual clarity. *Criminology and Public Policy*, 7, 203-213.

Ladd, H. F., Clotfelter, C. T., & Holbein, J. B. (2016). The growing segmentation of the charter school sector in north carolina. *Education Finance and Policy*.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education review*, *21*(1), 1–17.

LaGrange, R. L., Ferraro, K. F., & Supancic, M. (1992). Perceived risk and fear of Crime: Role of social and physical incivilities. *Journal of Research in Crime and Delinquency,* 29, 311-44.

Lefgren, L. (2004). Educational peer effects and the chicago public schools. *Journal of urban Economics*, *56*(2), 169–191.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review*, *29*(3), 480–488.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, *8*(4), 183–219.

Li, F., Zhang, C., & Feng, Q. (2014). An empirical analysis of relationship between teacher's quality and student's achievement. In *2014 international conference on education reform and modern management (ermm-14).*

Lincoln, R. D. (1997). Multi-year instruction: Establishing student-teacher relationships. *Schools in the Middle*, *6*(3), 50–52.

Lockwood, J., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439–467.

Lockwood, J., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ancova models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22–52.

Lockwood, J., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Lockwood, J., McCaffrey, D. F., et al. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, *1*, 223–252.

Manski, C. F., & Wise, D. A. (1983). *College choice in America*. Harvard University Press.

Martins, P. S., & Pereira, P. T. (2004). Does education reduce wage inequality? quantile regression evidence from 16 countries. *Labour economics*, *11*(3), 355–371.

Markowitz, F. E., Bellair, P. E., Liska, A. E. & Liu, J. (2001). Extending social disorganization theory: Modeling the relationships between cohesion, disorder, and fear. *Criminology*, 39, 293-319.

Matsueda, R. L. (1992). Reflected appraisals, parental labeling, and delinquency: Specifying as symbolic interactionist theory. *American Journal of Sociology*, 27, 1577–1611.

Maxfield, M. G., & Babbie, E. R. (2011). *Research Methods for Criminal Justice and Criminology* (6th ed.). Belmont: Wadsworth.

McCaffrey, D. F., & Lockwood, J. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, 773–797.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, *29*(1), 67–101.

McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* ERIC.

McCaffrey, D. F., Lockwood, J. R., Mihaly, K., Sass, T. R., et al. (2012). A review of stata commands for fixed-effects estimation in normal linear models. *Stata Journal*, *12*(3), 406.

McCaffrey, D. F., Sass, T. R., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

McGarrell, E. F., Giancomazzi, A. L., & Thurman, Q. C. (1997). Neighborhood disorder, integration, and the fear of crime. *Justice Quarterly*, 14, 479-500.

Mead, G. H. (1934). *Mind, Self, and Society*, (Charles W. Morris, ed.). Chicago: University of Chicago Press.

Mihaly, K., McCaffrey, D. F., Lockwood, J., Sass, T. R., et al. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. S*tata Journal*, *10*(1), 82.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. (2013). A composite estimator of effective teaching. *Seattle, WA: Bill & Melinda Gates Foundation*.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, *66*(4), 281–302.

Mincer, J. (1974). Schooling, experience, and earnings. human behavior & social institutions no. 2.

Monette, S., Bigras, M., & Guay, M.-C. (2011). The role of the executive functions in school achievement at the end of grade 1. J*ournal of Experimental Child Psychology*, *109*(2), 158-173. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022096511000191 doi: http://dx.doi.org/10.1016/j.jecp.2011.01.008

Morganstein, D., & Wasserstein, R. (2014). Asa statement on value-added models. *Statistics and Public Policy*, *1*(1), 108–110.

Murnane, R. J. (1975). The impact of school resources on the learning of inner city children.

Northeast, & at Brown University, I. R. E. L. (1997). *Looping: Supporting student learning through long-term relationships*. LAB, Northeast and Islands Regional Educational Laboratory, Brown University. Retrieved from https://books.google.com/books?id=JDYfnQEACAAJ

Ost, B., & Schiman, J. C. (2015). Grade-specific experience, grade reassignments, and teacher turnover. *Economics of Education Review*, *46*, 112–126.

Paufler, N. A., & Amrein-Beardsley, A. (2013). The random assignment of students into elementary classrooms implications for value-added analyses and interpretations. *American Educational Research Journal*, 0002831213508299.

Perkins, D. D. (1989). Description of data set and methods: NIMH final grant report appendix A. GRANTS #R01-MH40842-01 and -02. Unpublished Manuscript. Philadelphia, PA: Department of Criminal Justice, Temple University.

Perkins, D. D., Brown, B. B., and Taylor, R. B. (1996) The Ecology of empowerment: Predicting participation in community organizations. *Journal of Social Issues,* 52*,* 85-111.

Perkins, D. D., & Taylor, R. B. (1996). Ecological assessments of community disorder: Their relationship to fear of crime and theoretical implications. *American Journal of Community Psychology*, 24, 63-107.

Raudenbush, S., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis.* Thousand Oaks: Sage.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.

Rivkin, S. G. (1995). Black/white differences in schooling and employment. *Journal of Human Resources*, 826–852.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rivkin, S. G., & Schiman, J. C. (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal*, *125*(588), F425–F448.

Robinson, J. B., Lawton, B. A., Taylor, R. B., & Perkins, D. D. (2003). Multilevel longitudinal impacts of incivilities: Fear of crime, expected safety, and block satisfaction. Journal of *Quantitative Criminology*, 19, 237-74.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Rockoff, J. E., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics*.

Rodriguez, C., & Arenz, B. (2007). The effects of looping on perceived values and academic achievement. *ERS Spectrum*, *25*(3), 43–55.

Ross, C. E., & Mirowsky, J. (2001). Neighborhood disadvantage, disorder, and health. Journal of Health and Social Behavior, 42, 258-76.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rothstein, J. (2014). Teacher quality policy when supply matters. *The American Economic Review*, *105*(1), 100–130.

Rothstein, J. (2017). Revisiting the impacts of teachers. *Forthcoming: American Economic Review*.

Rothstein, J., & Wozny, N. (2013). Permanent income and the black-white test score gap. *Journal of Human Resources*, *48*(3), 510–544.

Rountree, P. W. (1998). A reexamination of the crime-fear linkage. *Journal of Research in Crime and Delinquency,* 35, 341-72.

Rountree, P. W., & Land, K. C. (1996a). Burglary victimization, perceptions of crime risk, and routine activities: A multilevel analysis across Seattle neighborhoods and census tracts. *Journal of Research in Crime and Delinquency*, 33, 147-80.

Rountree, P. W., & Land, K. C. (1996b). Perceived risk versus fear of crime: Empirical evidence of conceptually distinct reactions in survey data. *Social Forces*, 74, 1353-76.

Sampson, R. J. (2016). The characterological imperative: On heckman, humphries, and kautz's the myth of achievement tests: The ged and the role of character in american life. *Journal of Economic Literature*, *54*(2), 493–513.

Sampson, R. J., & Raudenbush, S. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology*, 105, 603-51.

Sampson, R. J., & Raudenbush, S. (2004). Seeing disorder: Neighborhood stigma and the social construction of "Broken Windows". *Social Psychology Quarterly*, 67, 319-42.

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, *38*, 9–23. *Schools and staffing survey.* (2017). Retrieved from https://nces.ed.gov/

Shadish, W. R., Cook, D. T. & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin Co.

Shamai, S., Ilatov, Z., Hertz-Lazarovitz, R., & Bentsvi-Mayer, S. (1995). Developing a model of classroom interaction between teacher and students. *Curriculum and Teaching*, *10*(2), 45–55.

Simel, D. (1998). Education for bildung: Teacher attitudes toward looping. *International Journal of Educational Reform*, *7*, 330–337.

Simonds, C. J., et al. (1994). What will happen if: Challenge behavior in the college classroom.

Skogan, W. G. (1990). *Decline and Disorder: Crime and the Spiral of Decay in American Neighborhoods*. New York: Free Press.

Steinberg, M. P., & Garrett, R. (2016a). Classroom composition and measured teacher performance. *Educational Evaluation and Policy Analysis*, *38*(2), 293-317. Re- trieved from http://dx.doi.org/10.3102/0162373715616249 doi: 10.3102/0162373715616249

Steinberg, M. P., & Garrett, R. (2016b). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317.

Taub, R. P., Taylor, D. G., & Dunham, J. D. (1984). *Paths of Neighborhood Change*. Chicago: University of Chicago Press.

Taylor, R. B. (1994). *Research Methods in Criminal Justice*. New York: McGraw Hill.

Taylor, R. B. (1996). Neighborhood responses to disorder and local attachments: The systemic model of attachment, social disorganization, and neighborhood use value. *Sociological Forum*, 11, 41-74.

Taylor, R. B. (1999). The incivilities thesis: Theory, measurement, and policy. In R. H. Langworthy (Ed.) *Measuring What Matters*. Washington, D.C.: National Institute of Justice/Office of Community Oriented Policing Services, 1999.

Taylor, R. B. (2001). *Breaking Away From Broken Windows: Baltimore Neighborhoods and the Nationwide Fight Against Crime, Grime, Fear and Decline*. New York: Westview.

Taylor, R. B. (2005). The incivilities or "Broken Windows" thesis. In L. E. Sullivan (Ed.) *Encyclopedia of Law Enforcement*. Thousand Oaks: Sage.

Taylor, R. B. (2010). Communities, crime, and reactions to crime multilevel models: Accomplishments and meta-challenges. *Journal of Quantitative Criminology*, 26, 455-66.

Taylor, R. B. (2012). Constructing Latent Variables for NIMH Common Data Set. (NIMH grants #R01-MH40842-01 and -02). Unpublished manuscript. Philadelphia, PA: Department of Criminal Justice, Temple University.

Taylor, R. B. & Covington, J. (1993). Community structural change and fear of crime. *Social Problems*, 40, 374-97.

Taylor, R. B., & Hale, M. (1986). Testing alternative models of fear of crime. *Journal of Criminal Law and Criminology*, 77, 151-89.

Thompson, N. L., Franz, D. P., & Miller, N. C. (2009). Research summary: Looping.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33.

Todd, P. E., & Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital*, *1*(1), 91–136.

Tracy, P. E., & Carkin, D. M. (2014). Adjusting for design effects in disproportionate stratified sampling designs through weighting. *Crime & Delinquency*, 60, 306-25.

U.S. Bureau of Economic Analysis. (2011). Downloaded April 15, 2012 from Baltimore Metropolitan Council: http://www.baltometro.org/content/view/148/210/#sources.

U.S. Census Bureau. (2009). Population, households, population by age, population by race, labor force and income. Data from the American Community Survey, 2009. Downloaded April 15, 2012 from Baltimore Metropolitan Council: http://www.baltometro.org/content/view/148/210/#sources.

Wallace, D. (2011). A test of the routine activities and neighborhood attachment explanations for bias in disorder perceptions. *Crime and Delinquency*, doi:10.1177/0011128711426538.

Wickes, R, Hipp, J. R., Zahnow, R., & Mazzerolle, L. (2013). 'Seeing' minorities and perceptions of disorder: Explicating the mediating and moderating mechanisms of social cohesion. *Criminology*, 51(3), 519-560.

Wilson, J. Q. (1975). *Thinking About Crime*. New York: Basic Books.

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly*, 211, 29-38.

Winters, M. A., Haight, R. C., Swaim, T. T., & Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, *34*, 69–75.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of econometrics*, *68*(1), 115–132.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *The American Economic Review*, *93*(2), 133–138.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach.* Nelson Education.

Worrall, J. L. (2006). Does targeting minor offenses reduce serious crime? A provisional, affirmative answer based on analysis of county-level data. *Police Quarterly*, 9, 47-72.

Wyant, B. R. (2008). Multilevel impacts of perceived incivilities and perceptions of crime risk on fear of crime: Isolating endogenous impacts. *Journal of Research in Crime and Delinquency*, 45, 39-64.

Xu, Y., Fiedler, M. L., & Flaming, K. H. (2005). Discovering the impact of community policing: The broken windows thesis, collective efficacy, and citizens' judgment. *Journal of Research in Crime and Delinquency, 42, 147-86.*