

QUALITY MATTERS: THE INFLUENCE OF TEACHER EVALUATION POLICIES AND SCHOOL
CONTEXT ON TEACHING QUALITY

By

Jihyun Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Educational Policy – Doctor of Philosophy

2017

ABSTRACT

QUALITY MATTERS: THE INFLUENCE OF TEACHER EVALUATION POLICIES AND SCHOOL CONTEXT ON TEACHING QUALITY

By

Jihyun Kim

Using a three paper format, this study examines how policy- and school-contexts might affect teaching quality. As many researchers have shown that teachers are one of the most important factors for student learning, accountability policies that used to target schools or districts now target individual teachers. That is, we are expecting more and more from our teachers. However, it is unclear that whether current policy- and school-context support teachers properly. This dissertation, which consists of three essays, examines this question with different perspectives.

The first essay investigates how teacher evaluation pressure perceived by early career teachers might affect their mathematics instruction. Drawing on observation and survey data, this essay examines whether current teacher evaluation policies conflict with ambitious mathematics instruction. As a result, early career teachers who felt a higher level of pressure related to teacher evaluation tended to be more active in using resources in their planning. However, those teachers were more likely to move further away from enacting ambitious mathematics instruction in terms of three dimensions in TRU Math scores, cognitive demand, agency, authority, and identity, and formative assessment. In other words, what teacher evaluation policies motivate teachers to do might not be aligned with ambitious mathematics instruction, and this is more salient with three dimensions of TRU Math. In terms of teacher individual level- (i.e., Mathematics Knowledge for Teaching) and school level- resources (i.e. school norms regarding mathematics instruction), only teachers' Mathematics Knowledge for Teaching (MKT) seems to have moderating effects on the association between evaluation pressure and teachers' enactment of mathematics instruction. A one unit increase in MKT made the negative effect of evaluation pressure almost doubled in two dimensions (i.e., agency, authority, and identity, and formative assessment). While social norms were expected to buffer such influence of teacher evaluation on teachers' instruction, there

was no significant moderating effect of social norms at schools. In total, it is arguable that teachers made rational decisions about their mathematics instruction, and teacher evaluation policies seemed to be prioritized while teacher level-and school-level resources did not buffer the influence of teacher evaluation policies.

The second essay is about the implementation and effects of teacher evaluation policies in Michigan school districts. Drawing on a loosely coupled system as a theoretical framework, this study examines whether Michigan school districts and the state government were loosely coupled in terms of teacher evaluations, what factors might have affected districts' decisions regarding teacher evaluations, and whether such policies produced any significant effects for student achievement. As a result, there was a clear variation in implementation of teacher evaluation policies, which showed loose coupling in the system: some school districts enacted the policies even before the state required them to do, while other school districts had never enacted the policies as the state mandated. The proportion of White students, fiscal resources available at the district level, student achievement, and leadership seemed to affect districts' decision-making related to teacher evaluation policies. Moreover, based on interrupted time series, the implementation of teacher evaluation policies had no significant effects on student achievement.

The third essay examines how principal leadership might affect early career teachers' turnover. Although teachers' instructional practices and student achievement are important, only if enough teachers are fully committed to their positions in each school can we consider teaching quality. I conceptualized principal leadership as three aspects: instructional leadership, leadership related to student behavior management, and leadership related to creating a supportive culture. I found that principal leadership was consistently important for early career teachers' turnover for their first five years. Among the three aspects of leadership, leadership related to creating a supportive culture had a significant and negative association with teachers' leaving the first school. That is, when a principal had strong leadership in terms of creating a supportive culture among teachers, early career teachers working for the school were less likely to leave their school.

Copyright by
JIHYUN KIM
2017

ACKNOWLEDGEMENTS

What a journey it was! Looking back on the last five years here in Michigan, it certainly has been the hardest, toughest, and most rewarding journey of my whole life. As an international student and former teacher without much experience in research, completing this work was not just about writing papers for me. English is a foreign language, as well as an academic one. As many international students say, there have been some moments where just breathing was hard. I needed to learn how to write, how to speak, and more importantly, how to think. I am so glad that I have a chance to say THANK YOU to all the great people whom I learned from during this journey.

I would like to express gratitude to my incredible advisors—Dr. Peter Youngs, Dr. Ken Frank, and Dr. Anne-Lise Halvorsen whose mentoring enabled me to pursue this work. Dr. Youngs, I remember so clearly the moment that I daringly asked you to be my advisor. I think that it was the best thing that I’ve done in the last 5 years in my life! I can’t find the words to express my appreciation for you. You have taught me not only about research, but also about your approaches to scholarship and life. I am especially appreciative for every moment that you kindly said, “It’s OK. Everything is alright.”

Dr. Frank, I deeply appreciate the trust and encouragement that you have shown me for years. I learned a lot from you, not just about theories and research methods, but also about the genuine joy of doing good “science.” You have shown me how to thrive as a researcher and how enjoyable it is to discover new findings. I will never forget our research team meetings where great research ideas originated and were polished. Thank you so much.

Dr. Halvorsen, you were the instructor for my first class here at MSU. I was so fortunate to have you in the first semester when everything was so new and confusing. Your dedication to students motivated me to read, write, and think more. You have also been a great role model for me as a good teacher and researcher as well as a strong and loving woman. I could always share my frank concerns and feelings with you. You always support me and it means so much to me. I can’t thank you enough for your support.

I also owe special thanks to Dr. Bob Floden and Dr. Min Sun for their thoughtful comments, and I was incredibly fortunate to have a chance to work with both of you. The conversations that I have had with you always enlightened me and helped me to grow. Thank you so much.

I want to thank the many people who made this work possible. I appreciate our SAMI team members and PIs for generously allowing me to use the data. Their efforts; including long distance driving, videotaping, rating lessons, creating and mailing surveys, sending out gift cards, asking teachers to complete the survey again, again, and again, cleaning and analyzing data, having a long and impassioned conversation about our ratings, and encouraging each other throughout this process; are deeply appreciated. I was so lucky to work with you all. I would also like to give special thanks to all the Michigan district administrators who completed my survey and allowed me to interview them. They were deeply motivated to improve students' lives and to contribute to the collective intelligence about policy implementation. As a researcher, I feel obligated to pay them back with solid research studies that can improve their districts. Thank you. Additionally, I would like to express my appreciation to Dr. Madeline Mavrogordato for her help with the recruiting process.

Pursuing a Ph.D. degree abroad entails so much sacrifice from family members and friends. I missed numerous weddings, birthdays, funeral, and family events over the last five years. In particular, I regret that couldn't make my grandpa's funeral in 2015. I bear in mind my grandpa's saying, "Don't feel small." I will never forget how you were such a strong and loving person. Thank you for your support, and I am so sorry, Grandpa. I really appreciate for my parents'—Dr. Lee and Dr. Kim— endless love, being healthy and happy, and calming me down whenever I am too concerned about what would never have happened. You encouraged me to be brave for many decisions in my life including pursuing a Ph.D. degree abroad. You taught me diligence and commitment. "Thank you" is too small to express my appreciation for you.

My brother, Hyunwoo Kim; I really like you and am sorry for not being available when you needed me most. I know you are doing great and you are already an outstanding journalist.

Last but not least: friends and colleagues! I deeply appreciate your love, emotional support, and the many discussions, meals and coffees we had together: CH, JJ, SL, IK, DH, JA, UJ, JL, SB, AH, IC, YL, YY, RX, TC, BC, MJ, BL, HP, EP, NE and SG. Without you guys, I would have not been here. Thank you so much.

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xiii
INTRODUCTION	1
REFERENCES	6
Essay 1: Two Conflicting Forces: How Early Career Teachers’ Perceptions of Pressure Associated with Teacher Evaluation Policies May Affect Their Mathematics Instruction.....	10
Literature Review.....	13
School Accountability Policies and Teacher Instruction	13
Teacher Evaluation Policy and Teacher Instruction	16
Theoretical Framework	19
Method	23
Data.....	23
Measures	25
ECTs’ perceived pressure related to teacher evaluation policies	25
ECTs’ enactment of mathematics instruction	25
ECTs’ and social network members’ planning of mathematics instruction	26
ECTs’ mathematics knowledge for teaching	26
ECTs’ and social network members’ enactment of mathematics instruction	26
Analytical Approach	28
Results.....	31
Discussion	40
NOTES.....	47
APPENDICES	49
Appendix A TRU Math Rubric	50
Appendix B Results Using Three Evaluation Pressure Items.....	51
REFERENCES	56
Essay 2: Teacher Evaluation Policies in a Loosely Coupled System: Their Implementation and Effects in Michigan School Districts.....	62
Theoretical Framework.....	67
Literature Review.....	69
Districts’ Decision Making and Implementation of Policies	69
The Effects of Teacher Evaluation Policies.....	72
Michigan Teacher Evaluation Policies.....	76
Method	78
Data.....	78
Measures	81
The implementation of teacher evaluation policies	81
Factors that might affect the implementation of the policies.....	82
Student achievement	82
Analytical Approach	82
Results.....	87
Variations in the Implementation of Teacher Evaluation Policies.....	87

Factors that Might Affect the Implementation of Teacher Evaluation Policies.....	88
Effects of Teacher Evaluation Policies on Student Achievement.....	94
Discussion	99
NOTES.....	106
APPENDICES	108
Appendix A Mich. Comp. Laws § 380.1249	109
Appendix B Interview Protocol for District Administrators.....	111
Appendix C Multicollinearity Check.....	113
Appendix D Time Trend Using CITS Model	114
Appendix E Results using Comparative Interrupted Time Series Models	116
Appendix F Results using CITS Model with 2012-13 School Year as Cut-Off Point	120
REFERENCES	124
Essay 3: It is About the Culture: Early Career Teacher Turnover and Principal Leadership	131
Factors That Affect Teacher Turnover.....	135
Theoretical Framework	139
Method	141
Data.....	141
Measures	142
ECTs leaving the school and leaving the profession	142
ECTs' perceptions about principal leadership	143
Control variables.....	144
Analytical Approach	146
Results.....	151
Descriptive univariate analysis	151
Discrete time survival analysis on ECTs leaving the school	158
Discrete time survival analysis on teachers leaving the profession	167
Discussion	175
NOTES.....	182
APPENDICES	184
Appendix A Survey Items Used in Analysis	185
Appendix B Correlation Between Weights and Principal Leadership and Control Variables	188
Appendix C Analysis Including Interaction Terms Between Weights and Principal Leadership Variables	190
Appendix D Analysis Including Interaction Terms Between Race/School Size/Ratio of Racially Minority Students and Principal Leadership Variables Without Weights	194
Appendix E Results Using Replicate Weights Instead of Teacher Clustered Errors.....	198
Appendix F The Results Using Untransformed Weights	200
Appendix G The Results Using No Weights	214
REFERENCES	228

LIST OF TABLES

Table 1. Background Information on Participating Districts	23
Table 2. Descriptive Statistics for Key Variables	28
Table 3. Potential Effects of Teachers’ Perceived Pressure Related to Teacher Evaluation on Their Use of Resources in Planning.....	32
Table 4. Potential Effects of Teachers’ Perceived Pressure Associated with Teacher Evaluation on Teachers’ Enactment of Mathematics Instruction: Main Effects and Heterogeneous Effects Based on Teachers’ MKT.....	36
Table 5. Potential Effects of Teachers’ Perceived Pressure Associated with Teacher Evaluation on Teachers’ Enactment of Mathematics Instruction: Heterogeneous Effects Based on Social Norms	38
Table 6. TRU Math Summary Rubric.....	50
Table 7. Potential Effects of Teachers’ Perceived Pressure Related to Teacher Evaluation on Their Use of Resources in Planning (Using three items)	51
Table 8. Potential Effects of Teachers’ Perceived Pressure Associated with Teacher Evaluation on Teachers’ Enactment of Mathematics Instruction: Main Effects and Heterogeneous Effects Based on Teachers’ MKT (Using three items)	52
Table 9. Potential Effects of Teachers’ Perceived Pressure Associated with Teacher Evaluation on Teachers’ Enactment of Mathematics Instruction: Heterogeneous effects Based on Social Norms (Using three items)	54
Table 10. Descriptive Statistics for School Districts	80
Table 11. The Timing of the Enactment of Teacher Evaluation Policies	88
Table 12. Factors that Might Affect the Implementation of Teacher Evaluation Policies	90
Table 13. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Mathematics)	96
Table 14. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Reading)	97
Table 15. Multicollinearity Check for Logistic Regressions	113
Table 16. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Mathematics)	116
Table 17. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Reading)	118

Table 18. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Mathematics)	120
Table 19. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Reading)	122
Table 20. Descriptive statistics	152
Table 21. Survivor Function	154
Table 22. Univariate Cox Regression Based Test for Quality of Survival Curves.....	157
Table 23. The Influence of Principal Leadership on Leaving the School.....	159
Table 24. The Heterogeneous Effects of Principal Leadership on Leaving the School.....	164
Table 25. The Influence of Principal Leadership on Leaving the Profession	168
Table 26. The Heterogeneous Effects of Principal Leadership on Leaving the Profession.....	172
Table 27. Survey items used in analysis	185
Table 28. Correlation Between Weights and Principal Leadership and Control Variables	188
Table 29. The Influence of Principal Leadership on Leaving the School (Including interaction terms) ..	190
Table 30. The Influence of Principal Leadership on Leaving the Profession (Including interaction terms)	192
Table 31. The Influence of Principal Leadership on Leaving the School Without Weights	194
Table 32. The Influence of Principal Leadership on Leaving the Profession Without Weights.....	196
Table 33. The Influence of Principal Leadership on ECT Leaving the School and Leaving the Profession (Using replicate weights)	198
Table 34. The Influence of Principal Leadership on Leaving the School (Using untransformed weights)	200
Table 35. The Heterogeneous Effects of Principal Leadership on Leaving the School (Using untransformed weights)	204
Table 36. The Influence of Principal Leadership on Leaving the Profession (Using untransformed weights)	207
Table 37. The Heterogeneous Effects of Principal Leadership on Leaving the Profession (Using untransformed weights)	211
Table 38. The Influence of Principal Leadership on Leaving the School (Without weights)	214

Table 39. The Heterogeneous Effects of Principal Leadership on Leaving the School (Without weights)	218
Table 40. The Influence of Principal Leadership on Leaving the Profession (Without weights)	221
Table 41. The Heterogeneous Effects of Principal Leadership on Leaving Profession (Without weights)	225

LIST OF FIGURES

Figure 1. Factors that Potentially Affect Teachers' Instruction.....	19
Figure 2. Time Trend in the Proportion of Proficient Students in Mathematics.....	94
Figure 3. Time Trend in the Proportion of Proficient Students in Reading	95
Figure 4. Time Trend in the Proportion of Proficient Students in Mathematics (CITS model)	114
Figure 5. Time Trend in the Proportion of Proficient Students in Reading (CITS model)	115
Figure 6. Time Line of BTLS Data Collection	142
Figure 7. Kaplan-Meier survival Curves for Teachers Leaving the School.....	155
Figure 8. Kaplan-Meier Survival Curves for Teacher Leaving the Profession.....	156

INTRODUCTION

As many researchers have shown, teachers are one of the most important factors for students' learning (Aaronson, Barrow, & Sander, 2007; Koedel & Betts, 2007; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain 2005; Rockoff, 2004; Sanders, Wright, & Horn, 1997). However, teachers are different from other physical inputs for students' learning, such as school facilities, small class size, or computers in classrooms, which are easily manipulable. Teachers themselves are human beings with their own motivations and characteristics as are students. More importantly, teachers are not working by themselves, but working within the school organizational context (e.g., Coburn, 2001; Coburn & Russell, 2008; Frank, Zhao, & Borman, 2004; Jackson & Bruegmann, 2009; Louis, Marks, & Kruse, 1996; Spillane, Kim, & Frank, 2012; Sun, Frank, Penuel, & Kim, 2013). Thus, it is not easy to change teachers and teaching quality despite their obvious importance.

Among various attempts to increase teaching quality, such as teachers' professional development, teacher preparation, improved curriculum, and teacher induction, teacher evaluation policies have emerged in recent years as one of the most popular tools (Delvaux et al., 2013; Hallinger, Heck, & Murphy, 2014; Measures of Effective Teaching Project (MET), 2013). The federal government has spurred this focus on teacher evaluation through Race to the Top and Title I ESEA (Elementary and Secondary Education Act) waivers (Ballou & Springer, 2015; Harris, Ingle, & Rutledge, 2014; Herlihy et al., 2014; Pogodzinski, Umpstead, & Witt, 2015; Steinberg & Sartain, 2015). Teacher evaluation policies aim to improve teaching quality by filtering out poor performers, giving feedback and support, and creating a results-oriented school culture (Hallinger, Heck, & Murphy, 2014). These mechanisms are designed to achieve the primary goal of improving teaching quality, with the ultimate goal of enhancing students' learning. The question here is whether the policies reach the primary and ultimate goals, given the attributes of teachers and school organizational conditions. Because of the short history of current teacher evaluation policies, there have been a limited number of research studies about the impacts of these policies based on empirical data. My first and second dissertation essays focus on this question from different perspectives.

The first dissertation essay addresses this question given the primary goal of the policies, improving teaching quality. The main research question for the first essay is “How are early career teachers’ perceptions of pressures associated with teacher evaluation policies related to their mathematics instruction?” To be specific, I focus on teachers’ planning and enactment of ambitious mathematics instruction, in order to examine if teacher evaluations encourage teachers to teach ambitious mathematics. Drawing on observation data and survey data collected during the 2015-16 school year as part of a larger study called the Study of Ambitious Mathematics Instruction, I found that teachers who perceived more pressure associated with teacher evaluation tended to move further away from enacting ambitious mathematics instruction. Interestingly, teachers who perceived a higher level of pressure associated with teacher evaluation were more likely to be active in using resources outside of the classroom for their lesson planning. Another important finding is that the association between teachers’ mathematical knowledge for teaching (MKT) and ambitious mathematic instruction became weaker when teachers perceived a high level of pressure associated with teacher evaluation policies. Social norms at each school did not affect the association between teacher evaluation pressure and teachers’ enactment of instruction.

These results indicate that current teacher evaluation policies might fail to motivate teachers to teach mathematics in ambitious ways. Between two conflicting forces, teacher evaluation and the demand for high-quality teaching, teachers need to make a choice, and it seems that teacher evaluation takes precedent over ambitious mathematics instruction. More importantly, neither individual teachers’ MKT nor school level social norms regarding mathematics instruction could buffer the effects of teacher evaluation. This study adds nuance to research on the implementation and effects of teacher evaluation policies as it focuses on ambitious mathematics and the role of resources (i.e., teachers’ MKT and social norms at the school), drawing on observation data, rather than depending only on teachers’ self-reports about their practice.

The second essay examines variation in the implementation of teacher evaluation policies in Michigan school districts; the factors might affect such variation; and the effects of such implementation. The unit of analysis for this essay is the district because each district determined the timing of policy

enactment and the details of the policies (e.g., components of teacher evaluation, weight for each component, and use of results). According to survey data collected from district administrators in 2015-16, there was indeed wide variation in the implementation of teacher evaluation policies. Some districts enacted the policies even earlier than the state required, while other districts had not enacted the policies as required as of 2015-16. This variation suggests that although the policies themselves represent a movement toward tight coupling, each district is still loosely coupled with the state government in terms of implementation of the policies. The proportion of White students, district total revenue, and prior student achievement had a significant association with districts' timing of the enactment of teacher evaluation policies. Among these factors, student achievement deserves a closer look; student achievement had a positive and significant association with districts not complying with the policies. This implies that districts did not perceive teacher evaluation policies as a promising tool for improving student achievement. According to the results about the effects of implementation of the policies at the district level reported in this essay, such districts' perceptions might be well grounded. Based on an interrupted time series (ITS) model, teacher evaluation policies had almost no influence on student achievement scores on Michigan Educational Assessment Program (MEAP) tests. Given the considerable amount of resources that each district needs to spend on implementing teacher evaluation policies, this result indicates that current teacher evaluation policies might be another administrative burden for districts rather than an effective tool for addressing the issue of low performance.

This second dissertation essay examines a similar question but from different perspectives and at a different level compared to the first essay. At the individual teacher level, the first essay addresses the primary goal of the policies, which is improving teaching quality. Focusing on the district level, the second essay evaluates the policies based on their ultimate goal, which is enhancing students' learning. Although the research sites for the two essays are different, combining the results from them leads to an important policy implication; current teacher evaluation policies do not seem to achieve their primary goal, i.e., improving teaching quality, or their ultimate goal, i.e., enhancing student achievement. While it is unclear whether it would be necessary to revise the policy design or provide more resources to support

the implementation process, it seems clear that current teacher evaluation policies should be changed in order to achieve their goals.

Teacher turnover is also an important matter for teaching quality. Only if enough teachers are fully committed to their positions in each school can we consider various ways to enhance teaching quality. Unfortunately, however, this might not be true for some schools that chronically suffer from a teacher shortage problem. It is not realistic to consider teaching quality in this context because simply hiring enough certified teachers is challenging for those schools. Moreover, early career teachers (ECTs), who have less experience and expertise, are more likely to fill vacancies in those schools (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008). A larger problem is that the teacher turnover rate is significantly higher among ECTs as compared to experienced teachers (Allensworth, Ponisciak, & Mazzeo, 2009). Taken together, schools with a severe teacher shortage problem, which mainly serve low-socio-economic (SES), low-achieving, and minority students, can have a severe teacher churning problem, causing many challenges with regard to school organization as well as students' achievement (Ronfeldt, Loeb, & Wyckoff, 2013). On the other hand, raising the retention rates of ECTs might improve students' learning not only in those school contexts, but other schools as well, given a significant and positive association between teachers' years of experience and students' achievement gain (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2011; Clotfelter, Ladd, & Vigdor, 2007; Henry, Bastian, & Fortner, 2011). Thus, retaining ECTs is as important as teacher evaluation for improving teaching quality. The question here is that in the situation where we cannot change student composition or invest a significant amount of resources to enhance ECTs' working conditions, how can we motivate teachers to stay in their school and/or the profession for a longer period of time?

The third dissertation essay examines this question, focusing on principal leadership as an important aspect of school context. Although several research studies have shown that principal leadership affects teachers' planned retention decisions as well as their actual turnover rates (Boyd, Grossman, Ing, Lankford, Loeb, & Wyckoff, 2011; Ingersoll & May, 2012; Ladd, 2011; Youngs, Kwak, & Pogodzinski, 2015), few studies have examined how different aspects of principal leadership shape

teacher turnover. I conceptualize principal leadership as featuring three related aspects: instructional leadership, leadership related to managing student behavior, and leadership related to creating supportive culture.

I draw on two nationally representative surveys, the Schools and Staffing Survey (SASS) collected in 2007-08 and the Beginning Teacher Longitudinal Survey (BTLS) collected from 2007-08 through 2011-12. I apply a discrete time survival analysis to take into account the longitudinal nature of the data. As a result, principal leadership had a consistent impact on whether ECTs left their school during the first five years. In particular, principal leadership related to creating supportive culture had a strong negative association with teachers leaving their first school. This result indicates that supporting ECTs is not only a job for formal leaders, but also for other teachers at the same school. In contrast, the association between principal leadership and ECTs leaving the profession was weak. ECT attrition from the profession was more closely related to the attributes of the occupation and teachers themselves, such as salary and teachers' commitment and perceptions about their preparation.

Teachers make rational decisions about their instructional practice and future career. Whether they are encouraged to teach in ambitious ways with a proper motivation system and to stay in their school and/or the profession for a longer period of time are essential issues for student learning. The following three essays illuminate how policy contexts and school organization can affect such decisions using different data, methods, and theoretical frameworks.

REFERENCES

REFERENCES

- 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.
- Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago public schools*. Chicago, IL: Consortium on Chicago School Research.
- Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance some problems in the design and implementation of evaluation systems. *Educational Researcher*, 44(2), 77–86.
- Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, 48(2), 303–333.
- Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, 27(4), 793-818.
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6), 673–682.
- Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, 23(2), 145–170.
- Coburn, C. E., & Russell, J. L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis*, 30(3), 203-235.
- Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching & Teacher Education*, 36, 1–11.
- Frank, K. A., Zhao, Y., Penuel, W. R., Ellefson, N., & Porter, S. (2011). Focus, fiddle and friends: Experiences that transform knowledge for the implementation of innovations. *Sociology of Education*, 84(2), 137-156.
- Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, 26(1), 1-24.
- 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.
- Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher*, 40(6), 271-280.
- Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation

- systems. *Teachers College Record*, 116(1), 1-28.
- Ingersoll, R. M., & May, H. (2012). The magnitude, destinations, and determinants of mathematics and science teacher turnover. *Educational Evaluation and Policy Analysis*, 34(4), 435–464.
- 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), 85-108.
- Koedel, C., & Betts, J. R. (2007). Re-examining the role of teacher quality in the educational production function. Working Paper 2007-03. *National Center on Performance Incentives*.
- Ladd, H. F. (2011). Teachers' perceptions of their working conditions how predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis*, 33(2), 235-261.
- Louis, K. S., Marks, H. M., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal*, 33(4), 757–798.
- Measures of Effective Teaching Project. (2013). *Ensuring fair and reliable measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237–257.
- Pogodzinski, B., Umpstead, R., & Witt, J. (2015). Teacher evaluation reform implementation and labor relations. *Journal of Education Policy*, 30(4), 540–561.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.
- Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2), 247–252.
- Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, 50(1), 4–36.
- Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11(1), 57–67.
- Spillane, J. P., Kim, C. M., & Frank, K. A. (2012). Instructional advice and information providing and receiving behavior in elementary schools: Exploring tie formation as a building block in social capital development. *American Educational Research Journal*, 49(6), 1112-1145.
- Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy*, 27(4), 793-818.
- Sun, M., Frank, K. A., Penuel, W. R., & Kim, C. M. (2013). How external institutions penetrate schools through formal and informal leaders. *Educational Administration Quarterly*, 49(4), 610-644.
- Youngs, P., Kwak, H. S., & Pogodzinski, B. (2015). How middle school principals can affect beginning

teachers' experiences. *Journal of School Leadership*, 25(1). 157-189.

Essay 1: Two Conflicting Forces: How Early Career Teachers' Perceptions of Pressure Associated with Teacher Evaluation Policies May Affect Their Mathematics Instruction

As “all students should learn and (that) learning should involve complex ideas and performance” (Lampert, Beasley, Ghouseini, Kazemi, & Franke, 2010, p.129) has become a mantra for educators and policy makers, “ambitious” mathematics instruction has been considered a gold standard for teachers. Moreover, this idea is emphasized by the Common Core State Standards (CCSS). The question here involves teachers’ motivation to teach mathematics in a more ambitious way; some in-service teachers need to change their instruction and, thus, without a proper system that motivates teachers to do this, ambitious instruction might only remain a slogan. On the other hand, teacher evaluation policies that provide teachers with clear motivation have drawn significant attention from both policy makers and researchers as an important tool for improving teaching quality. Under current teacher evaluation policies in most states, individual teachers are responsible for their students’ learning and multiple measures of teaching quality are used to determine their job status (Jiang, Spote, & Luppescu, 2015). In evaluation settings, teachers are likely to make efforts to achieve higher ratings or receive positive comments from their evaluators based on a certain evaluation tool, and in this process, teachers may be easily motivated to change their instruction. The question here is whether what current teacher evaluation policies prompt teachers to do is aligned with the idea of ambitious mathematics instruction.

According to utility theory, teachers, as important agents, make rational decisions about how they teach mathematics based on expected payoffs estimated by their own resources, social norms at the school level, and the broader policy context (Frank, Kim, & Belman, 2010). To be specific, a production function in this context might include the trade-off teachers can face between teaching high-quality mathematics versus teaching in a way that he/she can achieve a higher evaluation rating. The former is related to teachers’ self-efficacy and the latter is related to teacher evaluation pressure. However, the extent to which these two factors are incompatible in practice is unclear. In other words, can teachers pursue both together or do they need to choose one exclusively? When teachers need to make a choice between two, if the former one is chosen, the resources spent on teacher evaluation would be wasted, and if the latter one

is chosen, teacher evaluations will produce an unintended consequence, such as teachers teaching to standardized tests.

In order to examine how teacher evaluation policies affect teachers' ambitious mathematics teaching, I focus on teachers' planning and enactment of instruction as outcomes. Despite the significance of the topic, there is little known about how teachers experience these policies (Donaldson, 2012; Goldhaber, 2015; Harris & Herrington, 2015; Riordan, Lacireno-Paquet, Shakman, Bocala, & Chang, 2015; Youngs & Haslam, 2012). To address this gap, some researchers have surveyed teachers to learn about their perceptions of teacher evaluation policies (i.e., Delvaux et al., 2013; Donaldson, 2012; Jiang et al., 2015; Tuytens & Devos, 2009, 2010), and others have studied the impact of teacher evaluation on students' achievement scores without examining the underlying mechanisms by which evaluation affects instruction (i.e., Taylor & Tyler, 2012). Although these research studies provide important findings regarding policy implementation, there is still a missing piece: The possible influence of the policies on teachers' actual planning and enactment of instruction. Teachers can perceive these policies as important, legitimate, and useful for their improvement, but they may not change their instruction based on information from the evaluation process. In addition, teachers who increase their students' test scores after going through teacher evaluation may not have contributed to students' learning, but only reallocated time for teaching test-specific skills.

To operationalize the influence of teacher evaluation policies, I use teachers' perceived pressure associated with teacher evaluation policies as a proxy. To be sure, self-reported perceptions of "pressure" might not represent all of the ways that policies influence teachers; the influence of policies can be manifested in other ways. However, pressure in this context is more about the ways that policies lead teachers to change their instruction. The main question is the direction of such influence. If a teacher perceives a large amount of pressure to change their instruction because of teacher evaluation, how will he/she change their planning and enactment?

In addition, I investigate the role of resources in moderating the potential effects of teacher evaluation pressure on their instruction. Cohen, Raudenbush, and Ball (2003) argued that although

resources, including conventional resources (e.g., class size and textbooks) as well as resources in a broader sense (e.g., teachers' knowledge and school leadership) can be essential for instruction, "their value is likely to depend" on how they are used (p. 138). In this study, two types of resources are examined: at the individual teacher level, i.e., teachers' mathematical knowledge for teaching (MKT), and at the school level, i.e., other teacher colleagues' mathematics instruction. Teacher evaluations are likely to affect how teachers use these resources, since evaluations set a standard for performance and motivate teachers to change their instruction to conform to this standard. To be specific, teachers with a high level of MKT can use their resources for teaching mathematics in a way that enhances students' mathematical thinking without teacher evaluation pressure. However, it is also possible that they use this resource to teach their students test-specific skills, so that their students can achieve higher scores on standardized tests, when teachers encounter a high level of pressure related to teacher evaluation. Social norms that emphasize a high level of mathematics teaching may be a good resource for teachers for ambitious mathematics instruction, and it can buffer the effects of teacher evaluation pressure.

Arguably, the impact of the policies can be more salient for early career teachers (hereafter, ECTs). While current policies require all staff in schools, regardless of their tenure status, to be evaluated in general, most districts have differentiated evaluation procedures for ECTs and experienced teachers. ECTs are usually observed a greater number of times in a school year, they have more frequent conversations with evaluators, and their end-of-year evaluation can directly affect their job status for the following year (Waters district¹, 2015). That is, the design of the policies themselves expects that ECTs' instruction will change more than that of experienced teachers. Moreover, given that teachers' effects on student learning generally increase at a greater rate at the beginning of their careers compared to later in their careers (Rockoff, 2004), teachers at this stage are more likely to be receptive to external feedback on their instructional practice. In other words, while experienced teachers might have well-established instructional practices, which will not easily change based on others' evaluation of them, ECTs with lesser expertise may be easily encouraged to change their practice based on the teacher evaluation process. Taken together, teacher evaluation policies might be more influential for teachers who are in the early

stages of their development as professionals. Thus, studying ECTs' instruction in light of teacher evaluation policies can help us understand the potential effect of the policies on teachers at all experience levels.

This study draws on data from a larger study, the Study of Ambitious Mathematics Instruction (SAMI). Data collection for the project took place during the 2015-16 school year in multiple districts in Michigan, Indiana, and Illinois. The analysis draws on survey and observation data from ECTs and survey data from their egocentric social network members (i.e., their mentors and colleagues). The rest of this essay is organized as follows: The next section briefly reviews relevant literature about teachers' responses to teacher evaluation and school accountability policies. The third section introduces the theoretical framework, research questions, and hypotheses, and is followed by the method section. Then fifth section reports results and the last section provides discussion and implications.

Literature Review

Due to the short history of current teacher evaluation policies, there have been few studies on how they have been implemented (Donaldson, 2012; Goldhaber, 2015; Harris & Herrington, 2015; Riordan et al., 2015; Youngs & Haslam, 2012). Many studies have addressed the validity and reliability of various teacher evaluation tools, such as observation instruments, student surveys, student growth models, and value-added measures. However, the issue is not whether each measure is valid, but whether they produce the intended goal, improved student learning (Harris & Herrington, 2015).

Fortunately, school accountability policies, which share common elements with teacher evaluation policies, have been implemented long enough for researchers to study their influence on different aspects of schooling, including teachers' instruction. In this section, I first review the literature about the influence of school accountability policies on teachers' instruction, and then turn to teacher evaluation policies and how teachers have reacted to such policies.

School Accountability Policies and Teacher Instruction

School accountability policies, such as *No Child Left Behind* (NCLB), have significant similarities with teacher evaluation policies. Under NCLB, schools need to meet minimum achievement

benchmarks determined by others (i.e., Adequate Yearly Progress) and student achievement needs to be measured by state standardized test scores (Horn, Kane, & Wilson, 2015). This process involves sanctions for schools that fail to meet the requirements. When the focus of accountability policies shifts from the school to teachers, the expectations and pressures placed on teachers are similar to those previously placed on schools. Teachers need to demonstrate their performance with “objective” measures of student achievement, usually state-standardized tests, and other measures. If a teacher fails to achieve a certain goal, a sanction can be imposed. Given such similarities, I review the impact of school accountability policies on teachers’ instruction in order to develop the framework for the current study.

Although the effect of school accountability policies has been a controversial issue, most researchers agree that the policies have actually affected teachers’ instruction in a considerable way, compared with other types of educational reforms (Booher-Jennings, 2005; Hamilton et al., 2007; Horn et al., 2015; Reback, Rockoff, & Schwartz, 2014; Rouse, Hannaway, Goldhaber, & Figlio, 2013; White & Rosenbaum, 2008). While some studies focused on the general response of teachers to school accountability policies, such as reallocating resources to tested subjects (Reback, 2008), focusing on so-called “bubble kids” (Booher-Jennings, 2005), and even cheating (Jacob & Levitt, 2003), some researchers specifically focused on teachers’ instruction in their classrooms as a result of school accountability policies.

In case studies of teachers in three states, Hamilton et al. (2007) showed that most teachers reported a moderate or great deal of changes in their instruction due to state standardized tests, which are part of school accountability policies. Teachers in the case studies described instances of narrowing the curriculum in their mathematics and science lessons to activities that were addressed by standardized tests. Students in their classrooms were more likely to be assigned individualized, test-like tasks after enactment of school accountability policies. Using a nationally representative dataset, the Schools and Staffing Survey (SASS) and Early Childhood Longitudinal Program (ECLS) data, Reback and colleagues (2014) found some evidence of teachers narrowing the curriculum. The authors first found a significant positive effect on achievement of attending schools with a high probability of not meeting AYP. But the

heterogeneous effect on students at different points of the distribution made it unlikely that those low-achieving schools genuinely increased the learning of all of their students. The authors also found that teachers working at these schools were more likely to focus on students who were at the margins with regard to achievement, to move away from a whole-class instructional approach, to emphasize the topics and types of problems that state tests were likely to cover, to spend more time teaching content and to pursue more effective teaching strategies. The last two activities can be interpreted as positive outcomes of accountability, although the other changes that teachers reported potentially narrowed instructional practice in a way that focused only on knowledge and skills easily measured by standardized tests and on certain types of students while ignoring other learning goals and other students.

This point corresponds with a finding that increases in standardized test scores under school accountability policies mainly occurred on test items addressing students' basic skills (Jacob, 2005). At the same research site as Jacob (2005), White and Rosenbaum (2008) conducted a qualitative study about the impact of school accountability policies on teachers' behavior. They confirmed Jacob's (2005) finding and showed a larger impact of the policies on teachers' instructional practice. In interviews, teachers reported making significant efforts to teach test-taking skills, rather than to enhance students' thinking. The authors pointed out how school accountability policies had changed school culture in a way that valued teachers whose students earned proficient scores on standardized tests, rather than those who taught high-level thinking skills.

While such negative effects of school accountability policies on teachers' instruction have been well documented, the positive sides of the policies also have drawn considerable attention from policy makers and researchers. As Stecher (2002) anticipated, high-stakes testing can be beneficial for teachers in that it allows them to learn about their students' needs and their own strengths and weaknesses in systemic ways. It can also motivate teachers to work "harder" and "smarter." In fact, some research studies showed improvement in students' test scores as states implemented school accountability policies (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2004; Rouse et al., 2013). That is, these policies have achieved some positive effects on teachers' instruction. However, as Horn and

colleagues (2015) pointed out, the policy did not provide guidance for how teachers could work “harder” and “smarter.” The “how to” part is largely left to teachers and administrators in each school. Thus, teachers’ learning opportunity for enhancing their instructional practice was dependent on their own efforts and school contexts. That is, the policies failed to provide proper support for implementation of the policies; instead, they only focused on outcomes by design.

Taken together, research studies have shown the following: 1) School accountability policies indeed led to some changes in teachers’ instructional practice; 2) Based on data on students’ test scores, there is some evidence that such policies were beneficial for certain groups of students; and 3) Some studies pointed out unintended consequences for teachers’ instruction. Given the similarities between school accountability policies and teacher evaluation policies, these points are potentially applicable to the context of teacher evaluation policies. Teacher evaluation policies can influence teachers’ planning and enactment and analyzing only students’ test scores might not be sufficient to understand what actually happened in schools. In the following section, I turn to literature about teachers’ instructional practice and teacher evaluation policies.

Teacher Evaluation Policy and Teacher Instruction

As noted earlier, there have been a limited number of empirical studies about the influence of current teacher evaluation policies on teachers’ actual instructional practice. Much scholarly work has examined teachers’ general perceptions of teacher evaluation policies and factors that affect such perceptions (e.g., Delvaux et al., 2013; Donaldson, 2012; Donaldson & Papay, 2012; Geijssels, Sleegers, Berg, & Kelchtermans, 2001; Halverson, Kelley, & Kimball, 2004; Jiang et al., 2015; Kimball, 2002; Milanowski & Herbert, 2001; Tuytens & Devos, 2010). These research studies focused on teachers’ buy-in of the policies, assuming that teachers who perceive the policies as necessary and valid will implement them as intended.

As Donaldson (2012) showed, however, teacher buy-in does not necessarily mean that teachers change their instruction based on teacher evaluation policies. Drawing on interview data from a mid-size urban district, Donaldson (2012) reported that teachers agreed on the necessity of evaluation reform and

felt positive about setting their own goals and working toward such goals. Interestingly, teachers reported that they changed their planning based on teacher evaluation, while they noted that teacher evaluation policies did not impact their instruction.

If teachers' self-reports about the influence of the policies are accurate, this finding might indicate that teacher evaluation policies affect teachers' planning and enactment differently. In the planning and enactment of mathematics instruction, teachers utilize their knowledge and expertise in teaching different ways. Based on a framework developed by Salloum and colleagues (2016), teachers 1) set appropriate learning goals, 2) develop and/or modify tasks at appropriate levels of cognitive demand, and 3) anticipate students' thinking while they plan lessons. These planning activities happen usually ahead of time as teachers have time to think about the class carefully. In contrast, teachers do not have much time to carefully craft their responses to students while they enact instruction which includes teachers 1) engaging in instructional dialogue with students, 2) focusing dialogue on intended mathematical goals, and 3) correctly interpreting students' thinking (Salloum et al., 2016). According to Donaldson (2012), when teachers make careful decisions about their lessons ahead, they might consider teacher evaluation. While they are actually teaching students, however, they might not be able to do so. This might mean that it would take more time for teachers to internalize what evaluation encourages them to do. Building on this, the current study examines the potential influence of teacher evaluation on teachers' planning and enactment separately.

However, there are still some possibilities that teachers' perceptions are in part an inaccurate measure of the influence of the policies. Cognitive dissonance theory points out that while people generally pursue consistency (or "*consonance*") within their perceptions and between perceptions and actions, there are some exceptions, e.g., people behave differently from how they rationalize about the situation (i.e., "*dissonance*") (Festinger, 1962). As McLaughlin (1987) argued, moreover, not only do teachers' perceptions affect their instructional practice, but instructional practice can also change perceptions of policy. Thus, while it is clear that teachers' perceptions are one of the appropriate measures of teachers' responses to the policies, this measure can miss some aspects of teachers' behaviors, when

teachers behave in ways that are not consistent with their perceptions. Therefore, the current study focuses on teachers' actual practice by using classroom observation data, and survey data from questions directly asking their planning, rather than depending only on teachers' self-reported perceptions of the policies.

On the other hand, some studies have focused on the influence of school context on teachers' perceptions of teacher evaluation policies. Coggshall, Rasmussen, Colton, Milton, and Jacques (2012) suggested several conditions for supporting teachers' learning in teacher evaluation settings, including a culture of trust; well-supported and effective coaches, teacher leaders, and principals; and time for collaboration. That is, under these conditions, teachers might be more receptive to feedback about their instruction and they might enhance their instruction. Jiang and colleagues (2015) also pointed out that when teachers perceived a strong professional learning community in their schools, they were more likely to view teacher evaluation policies positively, which is possibly linked to teachers' actively changing their instruction based on the policies. School leadership also affects teachers' perceptions of teacher evaluation policies (Colby et al., 2002; Davis, Ellett, & Annunziata, 2002; Halverson et al., 2004; Jiang et al., 2015). Teachers' perceptions of their principals' instructional leadership and their sense of trust with principals had significant positive associations with teachers' acceptance of the policies (Delvaux et al., 2013; Jiang et al., 2015; Tuytens & Devos, 2010). Indeed, teachers were more aware of the quality of their evaluators than that of the evaluation rubrics (Firestone et al., 2013), and teachers' perceptions about principals' feedback had a significant impact on teachers' acceptance of the policies (Kimball, 2002). School principals also can "play essential roles in determining the meaning and values of teacher evaluation in schools, and how teacher evaluation can extend beyond its ritualistic traditions to improve teaching and learning" (Davis et al., 2002, p. 288).

In sum, the context of the schools where ECTs work, such as opportunities for interaction with colleagues, might influence their response to teacher evaluation policies. In particular, this study focuses on the possible influence of ECTs' social network members as contextual factors that moderate the

association between ECTs' perceived pressures associated with teacher evaluation and changes in their instructional practice.

Theoretical Framework

Building on Cohen and colleagues' (2003) framework of "instruction as interaction" and use of resources in instruction (p. 124), I first explore different factors that might affect teachers' planning and enactment of mathematics instruction, then consider the influence of teacher evaluation in relation to those factors. To be specific, I incorporate different types of resources available for teachers for planning and enactment and the influence of an external pressure, teacher evaluation policies, into the previous framework of Cohen and colleagues (2003). Figure 1 presents the theoretical framework for the current study.

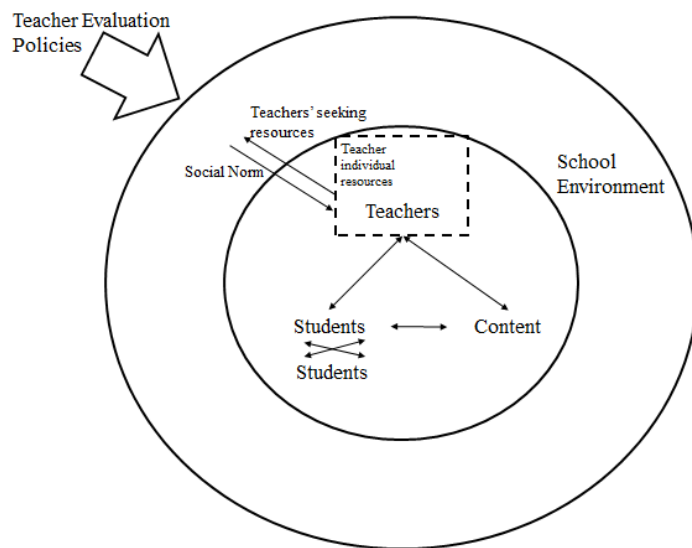


Figure 1. Factors that Potentially Affect Teachers' Instruction

First, teaching is not simply what teachers do, but it is an interactive process among teachers, students, and content (Cohen et al., 2003), and thus it is pertinent to consider mathematics instruction from all three aspects. From an ambitious mathematics instruction perspective, the questions are whether teachers' behavior is aligned with ambitious mathematics instruction, whether students learn how to think at a high mathematical level, and whether the content is selected and delivered in a way that supports students' high-level thinking. These accord well with the classroom observation tool that the SAMI

project used, TRU Math (Teaching for Robust Understanding in Mathematics) (Schoenfeld, Floden, & the Algebra Teaching Study and Mathematics Assessment Project, 2014). The purpose of this tool is to define and measure classroom interactions that enhance students' "robust understanding" of mathematical concepts (Schoenfeld, 2013, p. 608). The project used a modified version of the TRU Math rubric consisting of four dimensions: 1) the mathematics; 2) cognitive demand; 3) agency, authority, and identity; and 4) formative assessment [See Appendix A for The TRU Math Scoring Rubric]. These dimensions jointly capture all three aspects of instruction noted above.

The first dimension, the mathematics, captures the content side of instruction; the second dimension, cognitive demand, represents both content itself and how teachers deliver the content; the third dimension, agency, authority and identity, measures the extent to which students had an opportunity to engage in a high level of thinking; the fourth dimension, formative assessment, focuses on the interaction between teachers and students, and measures whether their discussion is helpful for students' high-level thinking. Thus, using TRU Math as a tool helps to capture teachers' enactment according to the framework of instruction as interaction. Moreover, since a lesson is coded based on four different dimensions, it is possible to examine the influence of teacher evaluation pressure on different aspects of teachers' instruction. For example, teacher evaluation pressure may affect the mathematics because this dimension is closely related to standardized tests. On the other hand, it would not affect agency, authority and identity because it is hard for a principal's quick walk-through or even standardized tests to capture this dimension.

Second, the school environment might affect mathematics instruction in various ways. In this study, I focus on two resources at different levels: social norms at the school level, and teachers' MKT at the individual teacher level. It has been widely reported that teachers' social network members have a significant influence on their instruction (Frank, Zhao, Penuel, Ellefson, & Porter, 2011; Sun, Penuel, Frank, Gallagher, & Youngs, 2013; Youngs, Frank, & Pogodzinski, 2012). Especially for ECTs, other teachers' planning and enactment, which contribute to school norms regarding mathematics instruction, are important resources. Although ECTs are equipped with fresh knowledge from their preparation

programs, they need to learn how to reconcile different demands from parents, their principal, and their district, and social norms at the school can show them how other experienced teachers have done this work. This function of social networks becomes particularly important when an external initiative, such as *NCLB* or reforms in curriculum, puts pressure on teachers to change their instruction (Coburn, 2001; Sun et al., 2013). Social norms at a given school regarding instruction can accelerate the penetration of external pressure on teachers, or they can filter out such influence. In the same vein, the influence of teacher evaluation on teachers' planning and enactment of mathematics instruction can differ across schools based on different social norms.

Teachers' MKT is also a key factor in the planning and enactment of mathematics instruction. MKT refers to "a kind of complex mathematical understanding, skill, and fluency used in the work of helping others learn mathematics" (Thames & Ball, 2010, p. 228). Criticizing the disconnect between subject matter courses in teacher preparation programs and what teachers actually do in mathematics classes, researchers have developed a tool to measure teachers' knowledge specific to teaching, rather than general knowledge possessed by most adults (Ball, Thames, & Phelps, 2008; Hill, Schilling, & Ball, 2004). MKT surveys have been shown to have a strong predictive validity; students taught by teachers with higher MKT scores had significantly higher gains (Hill, Rowan, & Ball, 2005); and teachers with higher MKT scores tended to teach rich mathematical content (Hill, Ball, Blunk, Goffney, & Rowan, 2007). However, having a high level of MKT does not guarantee that a teacher will engage in ambitious mathematics instruction; what MKT targets is not pedagogical quality itself, but knowledge that teachers need to teach mathematics in a solid way (Hill et al., 2007). Accordingly, I conceptualize MKT as a type of resource available at the individual teacher level. Teachers may or may not use this resource for teaching mathematics in ambitious ways, and the influence of teacher evaluation policies may depend on the use of this resource. In Figure 1, a dotted-line square surrounding teachers represents their MKT, as part of the available resources that can shape teachers' planning and enactment and moderate the influence of teacher evaluation pressure.

Third, teachers sometimes take an active role in seeking and using resources from their work environments during the planning process. They usually initiate this behavior due to their needs while social norms are likely to be established. It is important to study how teachers use resources from their environments because it illuminates the process of how teachers change their instruction, which classroom observation data do not necessarily capture. In other words, planning is a process for making changes in instruction and enactment is an outcome of this process; as noted above, teacher evaluation can affect these differently. Thus, it is important to investigate both in order to achieve a detailed picture of how teacher evaluation policies affect teachers' mathematics instruction. Therefore, I focus on changes in teachers' enactment of ambitious mathematics instruction and use of resources from their work environments in planning as the main outcome variables of interest.

Based on this theoretical framework, the research questions for this essay are as follows.

1. How are ECTs' perceptions of pressure associated with teacher evaluation policies related to their use of resources in planning?
 - 1a. How do early career teachers' MKT levels affect the association between their perceptions of pressure related to teacher evaluation policies and their use of resources in planning?
 - 1b. How do ECTs' social network members affect the association between their perceptions of pressure related to teacher evaluation policies and their use of resources in planning?
2. How are ECTs' perceptions of pressure associated with teacher evaluation policies related to their mathematics instruction?
 - 2a. How do early career teachers' MKT levels affect the association between their perceptions of pressure related to teacher evaluation policies and their mathematics instruction?
 - 2b. How do ECTs' social network members affect the association between their perceptions of pressure related to teacher evaluation policies and their mathematics instruction?

Method

Data

As part of a larger project, the SAMI project, this essay draws on data from surveys and classroom observations. In 2015-16, the SAMI study sampled all early career teachers in grades K-5 who had up to four years of full-time experience in elementary schools in eight districts in Michigan, Indiana, and Illinois. Among these three states, Michigan and Illinois implemented the CCSS and Indiana implemented a version of state mathematics standards that is very similar to the CCSS in 2015-16. All districts were small to medium-sized and served students from a range of socio-economic backgrounds.

Table 1. Background Information on Participating Districts

District*	State	The number of K-12 students	% of Free/Reduced lunch eligible students	% of White students
Ducasse	Michigan	19,000	70%	92%
Torres	Michigan	10,000	39%	91%
Garten	Illinois	6,000	5.3%	86%
Vongerichten	Illinois	17,000	63%	28%
Batali	Indiana	8,000	23%	79%
Henderson	Indiana	21,000	14%	77%
Lagasse	Indiana	15,000	61%	36%
Waters	Indiana	12,000	72%	32%

Note. * District names are pseudonyms. The number of K-12 students of each district is rounded for de-identification of school districts. Source: Common Core of Data (National Center for Education Statistics).

Table 1 reports background information for the sampled districts. Eligible ECTs were asked to complete three surveys, including an MKT survey and two surveys about their planning and enactment of mathematics lessons. They were also asked to participate in four observations of their mathematics lessons during the 2015-16 school year. The observations included both video ratings and live observations. In order to maintain a high level of inter-rater correlation (IRR), all raters had regular meetings to discuss the TRU Math observation rubric and specific cases (IRR: 0.505). The details of how the project selected the tool and trained raters are well documented in Salloum and colleagues (2016). A total of 500 ECTs were contacted for observations and surveys, and 84 of them completed all four observations and three surveys.

Most districts in the study had different teacher evaluation systems for tenured- and non-tenured- teachers, and the system for non-tenured teachers generally featured more frequent evaluations. In addition, the results from the evaluation of non-tenured teachers were used as a critical source of information for determining their employment status for the next school year. Although there was some variation in evaluation components across districts, most systems included student growth measures calculated at different levels (classroom-, building-, and/or district- level growth) and observation based on various tools, such as Five Dimensions of Teaching and Learning (Center for Educational Leadership, n.d.), Danielson's Framework for Teaching (Danielson, 1996), or a rubric developed by the district. Most districts used students' standardized tests scores for calculating student growth measures, but some districts deferred the student growth component altogether to the 2016-17 school year. Teachers were evaluated with regard to four or five levels of performance based on the weighted total score of each component of evaluation. Many districts in the study put the largest weights on observations conducted by building-level administrators.

In the first survey, administered in fall 2015, ECTs were asked to list their close teacher colleagues and formal mentor. Based on such nominations, nominated teachers were contacted and asked to complete two surveys, including an MKT survey and a survey about their planning and enactment of mathematics instruction.² The data collected from this process is egocentric data because we only focused on the social networks of certain people (i.e., ECTs). This contrasts with sociocentric network data, which includes data on the social networks of all teachers in a given school. We contacted 282 mentors and colleagues nominated by participating ECTs and asked them to complete two surveys in winter 2016. A total of 158 teachers completed both surveys (56% response rate). However, since many ECTs nominated another ECT working at the same school as their colleagues, some ECTs' data were included when social network exposure terms were calculated. The spring 2016 ECT survey also included questions about the respondents' close colleagues and mentors.

Measures

ECTs' perceived pressure related to teacher evaluation policies. In the spring 2016 ECT survey, items about teachers' perceptions of pressure associated with teacher evaluation policies were included. The items include: 1) The current teacher evaluation system has significantly affected my mathematics instruction; 2) I need to change my current teaching practices in order to earn a high score; 3) I need to earn a high teacher evaluation score to keep my job; and 4) I am concerned that my evaluation results can be used in making decisions ($\alpha=0.642$ with one significant factor)³. ECTs were asked about the extent to which they agreed with these statements. As noted above, these items measure how ECTs perceive the influence of policies on changes in their instruction. I took the mean of ECTs' responses to these four items and used it as the main independent variable for the analysis.

ECTs' enactment of mathematics instruction. The project conducted two back-to-back observations of each ECT in fall 2015 and again in spring 2016 to measure the quality of their mathematics teaching. The TRU Math observation rubric was used to assess their mathematics lessons with regard to four dimensions. Among the five dimensions of the TRU Math rubric, the SAMI project focuses on the following four dimensions: "The mathematics;" "cognitive demand;" "agency, authority, and identity;" and "formative assessment." Raters coded each lesson in 10-minute units, according to the rubric. The scale was from one to three treating non-mathematical activities as missing values and allowing .5 scales. For example, if a teacher taught a 90-minute class, a rater would code nine episodes according to the four dimensions. Thus, there would be 36 scores for this class ($9 \times 4 = 36$). It should be noted that the rubric is specified for each activity type, including whole class, individual work, and small group. Theoretically, the activity type should not affect ratings, but empirically, individual work episodes tended to be rated lower than other types and whole class instruction tended to be rated higher than other activity types.

The episode-level data for the four dimensions collected in spring 2016 was used as the main dependent variable for the analysis. Fall 2015 data is used as one of the main control variables in order to account for unobservables that may bias the analysis. For example, student characteristics and school

leadership can affect both teachers' perceived pressure associated with teacher evaluation and mathematics instruction. Since those factors might have already affected teachers' instruction in fall 2015, including this pre-measure helps reduce potential bias (Cook, Shadish, & Wong, 2008). The average scores in fall 2015 for each dimension were included at the teacher level as the pre-measure.

ECTs' and social network members' planning of mathematics instruction. In both the fall 2015 and spring 2016 surveys, a set of items about teachers' use of resources in planning was included. The stem asks, "In planning for the mathematics lessons that you taught during the past 2 weeks, how often did you make use of each of the following?" and the items include: 1) Your district mathematics pacing guide; 2) Advice from other teachers at your school; 3) Advice from your math instructional coach; 4) Performance criteria in teacher evaluation; and 5) Teacher evaluation results ($\alpha=0.719$ with one significant factor). Teachers answered "Never," "Sometimes," "Frequently," "Always," and "Not applicable (N/A)," with N/A coded as missing. I took the mean of teachers' responses to these items. The responses to the same set of items in the fall 2015 survey were used as pre-measures. The same set of items was used for calculating social network members' planning.

ECTs' mathematics knowledge for teaching. In order to measure the level of teachers' knowledge for teaching mathematics, a MKT survey (Hill et al., 2004) was administered to ECTs. This instrument measures teachers' knowledge for teaching various mathematical topics and different domains of teacher knowledge, such as knowledge of students and content. The MKT survey used for the SAMI project focused on elementary number and operation concepts and scores were provided as IRT scores by University of Michigan's Learning Mathematics for Teaching Project online system.

ECTs' and social network members' enactment of mathematics instruction. The surveys for ECTs and social network members included a set of questions about their enactment of mathematics instruction. The stem asks, "During the last 5 math lessons that you taught, in how many did your students have opportunities to do each of the following?" and the items include: 1) Verbally express their thinking; 2) Make connections between different strategies; and 3) Discuss other students' strategies. Teachers answered based on a scale of 0 lessons, 1-2 lessons, 3-4 lessons, and 5 lessons ($\alpha=0.791$ with

one significant factor). It should be noted that the main outcome variable for the analysis is ECTs' TRU Math scores; the survey-based measure of ECTs' enactment was only used when social network members' enactment was included in the analysis, in order to control for potential selection effects based on the ego's enactment.

In addition, various teacher-level and episode-level control variables were included in the analysis, in order to achieve estimates that are more robust. At the teacher level, whether a teacher taught a grade tested by state-standardized tests (i.e., 3rd- to 8th-graders for all three states of the current study), whether a teacher held an advanced degree, teacher gender, race, the number of students in the class, and years of experience working as a certified teacher were included. As noted above, however, the most important control variable is the teachers' pre-measures collected at the beginning of the 2015-16 school year. Since the analysis about teachers' enactment uses episode-level TRU Math scores, it is important to control for the attributes of each episode such as whether the observation was video or live, whether the lesson was the first day of a back-to-back observation, the type of episode (i.e., whole class, individual work, or small group), whether the episode was the first one, whether the episode was the last one, when the episode occurred in order (e.g., second episode or fifth episode), and the quadratic term of the order. Along with these control variables, district fixed-effects are included for all analyses in order to control for various attributes of districts that might affect both teacher evaluation pressure and enactment and planning, such as curriculum and student composition. Table 2 reports the descriptive statistics of the main variables for the analysis.

Table 2. Descriptive Statistics for Key Variables

Variable	M	SD	N
Episode level characteristics in spring 2016			
The mathematics	1.932	0.561	1172
Cognitive demand	1.716	0.579	1164
Agency, authority, and identity	1.433	0.514	1152
Formative assessment	1.469	0.558	1155
Types of episode-Whole class	0.387	-	1278
Types of episode-Individual work	0.235	-	1278
Types of episode-Small group	0.366	-	1278
Number of episodes per lesson	7.002	1.882	1278
Video observation	0.600	-	1278
Episode level characteristics in fall 2015			
The mathematics	1.915	0.415	1266
Cognitive demand	1.671	0.422	1266
Agency, authority, and identity	1.405	0.341	1266
Formative assessment	1.490	0.392	1266
ECTs' characteristics			
Evaluation pressure	2.124	0.547	95
Mathematical knowledge for teaching	-0.063	0.914	99
Use of resources for planning in fall 2015	1.754	0.623	95
Use of resources for planning in spring 2016	1.746	0.562	95
Enactment of instruction fall 2015	2.301	0.674	93
Enactment of instruction spring 2016	2.337	0.675	93
Teaching tested grade	0.370	-	100
Holding advanced degree	0.255	0.438	94
Total years of experience working as a certified teacher	2.911	1.421	90
Male	0.053	-	94
White	0.903	-	93
The number of students in the class	24.970	5.042	99
Social network members' characteristics			
Social network members' MKT	0.037	0.668	79
Social network members' use of resources for planning	1.814	0.412	89
Social network members' enactment of instruction	2.387	0.448	90

Note. Types of episode, teaching tested-grade, holding an advanced degree, male, and white are dummy variables. Average TRU Math score in fall 2015 is an average score across all episodes and all dimensions. Social network members' characteristics were calculated by taking weighted mean of nominated teachers' responses to the survey items by the frequency of interactions for each ECT.

Analytical Approach

The main goal of the analysis is to examine potential effects of teacher evaluation pressure on ECTs' planning and enactment of mathematics instruction from an ambitious mathematics instruction perspective. Before conducting the main analysis, it was important to determine the level of analysis for

each research question. Based on the research design, there can be four different levels: episode level TRU Math scores are nested in teachers, teachers are nested in schools, and schools are nested in districts. In order to take into account of the nested structure of the data, three- or four- level hierarchical linear modeling (HLM) may be considered theoretically (Raudenbush & Bryk, 2002). However, the school level was excluded because there were not enough teachers per school; there were on average 1.94 ECTs per school in the data. In addition, since there were only eight school districts in the data, district-fixed effects were included. Taken together, when episode-level data were available, I applied two-level HLM, and when only teacher-level data were available, I applied OLS regression with district fixed effects.

For the first research question about the association between evaluation pressure and teachers' use of resources in planning, only teacher-level data were available. The main model for the first research question is as follows:

$$Y_{ijt} = \alpha + \rho Y_{ijt-1} + \beta(\text{EvaluationPressure}_{ijt}) + \mathbf{X}_{ijt} \gamma + \mathbf{D}_{jt} + e_{ijt} \quad (1)$$

Where Y_{ijt} indicates resource use in planning of teacher i , in district j , at time t . Y_{ijt-1} represents the pre-measure of Y_{ijt} . $\text{EvaluationPressure}_{ijt}$ is teachers' perceived pressure associated with teacher evaluation. $\mathbf{X}_{ijt}$ is a vector of teacher i 's characteristics, including gender, race, years of teaching, holding a Master's degree or higher, number of students in the class, mathematical knowledge for teaching, and teaching a tested grade. $\mathbf{D}_{jt}$ is district fixed effects and e_{ijt} is a random error term. For the second question about moderating the effects of ECTs' MKT, a term for the interaction between ECTs' MKT and evaluation pressure is added to model (1).

$$Y_{ijt} = \alpha + \rho Y_{ijt-1} + \beta_{1j}(\text{EvaluationPressure}_{ijt}) + \beta_{2j}(\text{MKT} * \text{EvaluationPressure}_{ijt}) + \mathbf{X}_{ijt} \gamma + \mathbf{D}_{jt} + e_{ijt} \quad (2)$$

The next question is related to moderating effects of social norms related to mathematics instruction. Following the approach of Penuel, Frank, Sun, Kim, and Singleton (2013), social network exposure was calculated by taking the mean of the attributes of nominated mentors/close colleagues weighted by the frequency of interaction between the ECT and these individuals. This type of social network influence is more likely to be the norm, rather than information seeking, so the mean of those attributes, rather than their sum, needs to be used. Two attributes of social network members are

considered: social network members' use of resources in planning, and enactment of mathematics instruction. The social network exposure terms are specified as follows (Sun et al., 2013):

$$\text{SocialNorms}_i = 1/n_i \sum_{i'=1}^n (\text{Interaction}_{ii'}) X(\text{Colleagues' practice}_{i'}) \quad (3)$$

Where n_i is the total number of nominations that teacher i made in the fall and spring surveys. After creating two versions of this term, one for social norms about planning and one for enactment, the terms for interactions between these terms and evaluation pressure are included to examine whether the association between evaluation pressure and the outcome differs by social norms.

$$Y_{ijt} = \alpha + \rho Y_{ijt-1} + \beta_{1j}(\text{EvaluationPressure}_{ijt}) + \beta_{2j}(\text{SocialNorms} * \text{EvaluationPressure}_{ijt}) + \mathbf{X}_{ijt} \gamma + \mathbf{D}_{jt} + e_{ijt} \quad (4)$$

Research questions 2, 2a, and 2b involve TRU Math scores, which are at the episode level. Thus, a two-level HLM was applied; the first was the episode level and the second was the teacher level. The main model for these research questions is specified as follows:

$$\text{Level 1 (Episode level): } Y_{ijt} = \beta_{0j} + \mathbf{X}_{ijt} \delta + e_{ijt}$$

$$\text{Level 2 (Teacher level): } \beta_{0j} = \gamma_{00} + \rho Y_{jt-1} + \gamma_{01}(\text{EvaluationPressure}_j) + \mathbf{Z}_{jt} \lambda + \mathbf{D}_{jt} + u_{0j} \quad (5)$$

Where Y_{ijt} indicates a TRU Math score of episode i of teacher j , at time t . Scores of four dimensions are included separately. $\mathbf{X}_{ijt}$ is a vector of episode attributes: whether the observation was video or live, whether the lesson was the first day of a back-to-back observation, the type of episode (i.e., whole class, individual work, or small group), whether the episode was the first one, whether the episode was the last one, the episode's place in the order, and the quadratic term of the order. Y_{jt-1} is an average score of each dimension in fall 2015. For example, when the outcome is the dimension 1 (The mathematics) score of an episode in spring observations, Y_{jt-1} is the average score of dimension 1 in fall observations across all episodes, so it is included at the teacher level. $\mathbf{Z}_{jt}$ represents a vector of teacher characteristics noted above. $\mathbf{D}_{jt}$ is district-fixed effects, and e_{ijt} and u_{0j} is episode specific- and teacher specific-residuals, respectively. In order to estimate the moderating effects of ECTs' MKT and social norms on the association between evaluation pressure and teachers' enactment of mathematics instruction, the terms for interactions between 1) MKT and evaluation pressure, and 2) social norms and evaluation pressure were added to model (5) at the teacher level separately, in the same way as models (2) and (4).

Before creating the interaction terms, the variables were grand-mean centered. The models are specified as follows.

$$\text{Level 1: } Y_{ijt} = \beta_{0j} + \mathbf{X}_{ijt}\delta + e_{ijt}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \rho Y_{jt-1} + \gamma_{01}(\text{EvaluationPressure}_j) + \gamma_{02}(\text{MKT} * \text{EvaluationPressure}_j) + \mathbf{Z}_{jt}\lambda + \mathbf{D}_{jt} + u_{0j} \quad (6)$$

$$\text{Level 1: } Y_{ijt} = \beta_{0j} + \mathbf{X}_{ijt}\delta + e_{ijt}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \rho Y_{jt-1} + \gamma_{01}(\text{EvaluationPressure}_j) + \gamma_{02}(\text{SocialNorms} * \text{EvaluationPressure}_j) + \mathbf{Z}_{jt}\lambda + \mathbf{D}_{jt} + u_{0j} \quad (7)$$

Results

Before analyzing the main research questions, I started by examining ECTs' perceived pressure associated with teacher evaluation. First, I ran an unconditional model of HLM on teacher evaluation pressure to analyze whether evaluation pressure is an individual teacher-level and/or district-level phenomenon. Since districts have slightly different systems for evaluation, there may be some variance in evaluation pressure between districts. As mentioned above, school-level analysis is not feasible with the limited number of teachers per school. The Inter Class Correlation (ICC) was almost zero at the district level, so evaluation pressure seems to be an individual teacher-level phenomenon. This also supports using district fixed effects rather than including district level as another level in the HLM analysis. Then the question became which teachers felt more pressure. As an exploratory analysis, I examined a simple pair-wise correlation between teachers' perceived pressure and their characteristics, such as gender, race, teaching grades, math knowledge for teaching, whether they held an advanced degree, and their years of teaching. While most teacher characteristics did not have a significant association with evaluation pressure, White teachers perceived significantly less pressure ($\gamma = -0.292$, $p \leq 0.01$), and teachers who taught grades tested by a state-level standardized test (i.e., 3rd- to 8th-graders) perceived much more pressure ($\gamma = 0.195$, $p = 0.061$).

Next, I proceeded with the analysis for the research questions about the potential effects of teachers' perceived pressure related to teacher evaluation on changes in their use of resources in planning. Table 3 reports the results.

Table 3. Potential Effects of Teachers' Perceived Pressure Related to Teacher Evaluation on Their Use of Resources in Planning

	Model (1)	Model (2)	Model (3)	Model (4)
Use of resource in fall 2015	0.546*** (0.088)	0.545*** (0.087)	0.580*** (0.010)	0.514** (0.093)
Evaluation pressure	0.392** (0.104)	0.402*** (0.086)	0.878** (0.260)	0.559 (0.889)
MKT	-0.232*** (0.020)	-0.338 (0.388)	-0.167* (0.034)	-0.195* (0.038)
MKT*Evaluation pressure		0.107 (0.176)		
Holding an advanced degree	0.112 (0.068)	0.112 (0.064)	0.111 (0.110)	0.178 (0.148)
Total years in teaching	-0.133* (0.023)	-0.131 (0.027)	-0.211** (0.024)	-0.178** (0.018)
Male	0.014 (0.181)	0.010 (0.187)	0.035 (0.236)	0.020 (0.308)
White	-0.077 (0.088)	-0.078 (0.090)	-0.120 (0.101)	-0.153* (0.112)
Teaching a tested-grade	-0.074 (0.122)	-0.072 (0.124)	-0.042 (0.051)	-0.009 (0.106)
The number of students in class	-0.039 (0.018)	-0.041 (0.017)	-0.130 (0.009)	-0.132 (0.013)
Social norm regarding planning			0.526 (0.320)	
Social norm regarding enactment				0.156 (0.753)
Social norm regarding planning*Evaluation pressure			-0.690 (0.133)	
Social norm regarding enactment*				-0.197 (0.327)
Evaluation pressure				
Enactment in fall 2015				0.072 (0.055)
R-squared	0.634	0.634	0.660	0.659
N	79	79	74	73

Note. All models included district fixed effects. All coefficients are standardized and clustered robust standard errors at district level are in parentheses. Enactment in fall 2015 is included in model 4 to control for potential selection effect by ECTs' enactment level. * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

All coefficients are standardized and standard errors are clustered at the district level. In model 1, teachers' perceived pressure regarding teacher evaluation had a significant positive association with teachers' use of resources in planning after controlling for the pre-measure, teacher-level characteristics, and district fixed effects. That is, when a teacher felt more pressure related to evaluation, she/he was more likely to

frequently use resources outside of the classroom, such as a district pacing book, advice from other teachers, and teacher evaluation criteria and results, in the planning process. The magnitude of this association was also substantial. A one standard deviation increase in evaluation pressure was associated with a 0.392 standard deviation increase in resource use. Another interesting finding from this model was that teachers' MKT has a negative and significant association with teachers' use of resources in planning; teachers with high MKT scores tended to use resources less frequently. Even when I excluded evaluation pressure from the model, this association remained significant.

In model 2, I included a term for the interaction between teachers' MKT and evaluation pressure. This interaction term is far from the level of significance. In other words, teachers' MKT level did not seem to cause any changes in the association between teachers' perceived pressure and use of resources in planning. Models 3 and 4 included terms for the interactions between social norms regarding planning (model 3) and enactment (model 4) and evaluation pressure. Although both interaction terms were not significant at $\alpha=0.05$ level, the interaction term regarding social network members' planning was very close to significance ($p=0.051$). When a teacher's social network members actively used resources in their planning, the positive association between evaluation pressure and a teacher's use of resources in planning became weaker.

Based on the results, it is arguable that teacher evaluation motivates teachers to more actively use resources for their planning. However, whether this helps student learning is still questionable; it depends on how teachers use these resources (Cohen et al., 2003). Although the outcome of this planning behavior cannot be fully analyzed in the current study due to limits of the data, I explored the association between TRU Math scores and teachers' use of resources in planning. The correlation between these two variables is close to zero, and it was unchanged even when I ran an OLS regression controlling for other teacher-level covariates used for the main models. That is, although teacher evaluation may lead teachers to use more resources, it might not be linked to their enactment of ambitious mathematics instruction. I revisit this point in the discussion section.

Tables 4 and 5 report the results of the potential effects of teachers' perceived pressure associated with teacher evaluation on changes in teachers' enactment of mathematics instruction, including main effects and heterogeneous effects based on teachers' MKT levels and social norms. In these models, variables were not standardized because the outcome variable, raw TRU Math scores, is meaningful and interpretable based on the rubric. In terms of the main effects, teacher evaluation pressure had a significant and negative association with changes in teachers' ambitious mathematics instruction measured by dimensions 2 (cognitive demand), 3 (agency, authority, and identity), and 4 (formative assessment). There was no significant association between teachers' dimension 1 ratings (the mathematics) and evaluation pressure. That is, when ECTs perceived more pressure associated with teacher evaluation, they were more likely to move away from enacting ambitious mathematics instruction.

Such ECTs tended to teach less cognitively demanding tasks; students in their class tended to have fewer opportunities to talk about their ideas; and these ECTs were less likely to monitor student ideas and use them in the class. In contrast, the mathematical content they covered during the lesson was not affected by their perceived evaluation pressure. The magnitude of association was stronger for dimension 2 scores; a one-unit increase in teachers' perceived pressure was associated with a 0.146-point lower TRU math score. Given the considerable difference in 1 point based on the rubric (See Appendix A for details), this value is not negligible. The magnitude of the associations between evaluation pressure and dimensions 3 and 4 are similar but slightly less than those between evaluation pressure and dimension 2 ratings ($b=-0.104$, $p\leq 0.05$ and $b=-0.133$, $p\leq 0.05$, respectively).

Models 5 through 8 in Table 4 report the potential moderating effect of teachers' MKT. Those variables are centered for ease of interpretation and to account for multicollinearity. The term for the interaction between teachers' MKT and evaluation pressure had a negative and significant association with changes in teachers' TRU Math scores for dimensions 3 and 4. For teachers with an average level of evaluation pressure, a one-unit increase in their MKT score is associated with an almost doubled decrease in their enactment of ambitious mathematics instruction in terms of dimensions 3 and 4. This is quite a strong association; the coefficients of the main effect (i.e., evaluation pressure) are similar to those of the

interaction terms. On the other hand, teachers' MKT does not affect the association between evaluation pressure and TRU Math scores for the other two dimensions.

Table 5 presents the results regarding the potential heterogeneous effects of evaluation pressure according to social norms. In this analysis, I examined two different types of social norms at each school: social norms regarding planning and enactment. In all models, social norms at a given school did not have any influence on the association between evaluation pressure and changes in teachers' ambitious mathematics instruction. As a secondary analysis, I examined whether other characteristics of teachers, such as years in teaching, hours they have spent on professional development that addressed teacher evaluation, and their prior TRU Math scores, had any influence on the association between evaluation pressure and changes in teachers' ambitious mathematics instruction. For example, it is possible that more years of teaching in schools make the influence of teacher evaluation pressure weaker due to teachers having more expertise in mathematics instruction. Similarly, professional development hours might have enabled teachers to learn about their evaluation system more accurately, and could have changed the magnitude of the association between evaluation pressure and their mathematics instruction. However, the terms for interactions between those variables and evaluation pressure were not significant. In other words, the association between evaluation pressure and changes in teachers' TRU Math scores seems to be consistent for teachers, but only their MKT levels moderate this association.

Table 4. Potential Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation on Teachers' Enactment of Mathematics Instruction: Main Effects and Heterogeneous Effects Based on Teachers' MKT

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Formative assessment	Model (5) Dimension1: The mathematics	Model (6) Dimension2: Cognitive demand	Model (7) Dimension3: Agency, authority, and identity	Model (8) Dimension4: Formative assessment
Mean score in fall 2015	0.130 (0.094)	0.077 (0.087)	0.134 (0.087)	0.143 (0.083)	0.130 (0.094)	0.077 (0.086)	0.153 (0.086)	0.175* (0.080)
Evaluation pressure	-0.050 (0.067)	-0.146* (0.063)	-0.104* (0.050)	-0.133* (0.053)	-0.072 (0.070)	-0.174** (0.065)	-0.134** (0.051)	-0.181*** (0.052)
MKT	-0.070 (0.041)	-0.046 (0.039)	0.004 (0.031)	-0.066* (0.033)	-0.073 (0.041)	-0.050 (0.039)	0.0003 (0.031)	-0.074* (0.031)
Holding an advanced degree	-0.166 (0.091)	-0.142 (0.087)	-0.120 (0.069)	-0.123 (0.072)	-0.168 (0.090)	-0.145 (0.086)	-0.123 (0.068)	-0.129 (0.069)
Total years in teaching	0.018 (0.025)	-0.016 (0.024)	-0.013 (0.019)	-0.019 (0.020)	0.013 (0.025)	-0.021 (0.024)	-0.019 (0.019)	-0.028 (0.019)
Video observation	-0.316** (0.106)	-0.279** (0.102)	-0.014 (0.082)	-0.158 (0.086)	-0.335** (0.107)	-0.304** (0.102)	-0.040 (0.081)	-0.199* (0.082)
Teaching a tested grade	0.045 (0.084)	0.075 (0.080)	0.003 (0.064)	0.020 (0.067)	0.040 (0.084)	0.068 (0.080)	-0.003 (0.063)	0.010 (0.063)
The number of students in class	-0.002 (0.010)	0.0001 (0.010)	0.017* (0.008)	0.019* (0.008)	-0.001 (0.010)	0.001 (0.009)	0.018* (0.008)	0.020* (0.008)
Male	-0.344* (0.147)	-0.287* (0.138)	-0.287** (0.111)	-0.363** (0.115)	-0.325* (0.147)	-0.263 (0.138)	-0.258* (0.109)	-0.314** (0.111)
White	0.286* (0.129)	0.186 (0.123)	0.0329 (0.0992)	0.144 (0.103)	0.302* (0.129)	0.206 (0.123)	0.055 (0.098)	0.182 (0.099)
First day of back-to-back observation	0.079** (0.029)	0.083** (0.032)	0.054 (0.029)	0.066* (0.030)	0.078** (0.029)	0.083** (0.032)	0.054 (0.029)	0.065* (0.030)
Type: Individual work	-0.246*** (0.044)	0.014 (0.048)	-0.153*** (0.043)	-0.159*** (0.046)	-0.246*** (0.044)	0.013 (0.048)	-0.153*** (0.043)	-0.160*** (0.045)
Type: Small group	0.062 (0.041)	0.190*** (0.044)	0.038 (0.039)	0.110** (0.042)	0.062 (0.040)	0.190*** (0.044)	0.037 (0.039)	0.107** (0.041)

Table 4 (cont'd)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
First episode	-0.038 (0.059)	-0.083 (0.065)	-0.109 (0.058)	-0.115 (0.062)	-0.037 (0.059)	-0.082 (0.065)	-0.108 (0.058)	-0.113 (0.062)
Last episode	-0.206*** (0.054)	-0.186** (0.059)	-0.180*** (0.054)	-0.136* (0.057)	-0.207*** (0.054)	-0.188** (0.059)	-0.184*** (0.054)	-0.144* (0.057)
Episode order	-0.047 (0.027)	-0.050 (0.030)	-0.073** (0.027)	-0.076** (0.028)	-0.047 (0.027)	-0.049 (0.030)	-0.072** (0.027)	-0.075** (0.028)
Quadratic term of order	0.003 (0.002)	0.002 (0.003)	0.005* (0.002)	0.004 (0.002)	0.003 (0.002)	0.002 (0.002)	0.005* (0.002)	0.004 (0.002)
MKT*Evaluation pressure					-0.095 (0.088)	-0.120 (0.084)	-0.134* (0.066)	-0.219** (0.067)
N	986	978	972	976	986	978	972	976
N of teachers	84	84	84	84	84	84	84	84

Note. Types of episode, video observation, first day of back-to-back observation, first and last episode, teaching tested grade, holding an advanced degree, male, and white are dummy variables. All models included district fixed effects. Coefficients are not standardized because the unit of dependent variable, TRU Math score, is meaningful. In models 5 through 8, teachers' MKT and evaluation pressure are centered. * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 5. Potential Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation on Teachers' Enactment of Mathematics Instruction: Heterogeneous Effects Based on Social Norms

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Formative assessment	Model (5) Dimension1: The mathematics	Model (6) Dimension2: Cognitive demand	Model (7) Dimension3: Agency, authority, and identity	Model (8) Dimension4: Formative assessment
Mean score in fall 2015	0.194* (0.092)	0.098 (0.087)	0.261** (0.098)	0.262** (0.098)	0.176* (0.090)	0.086 (0.089)	0.259** (0.098)	0.258** (0.098)
Evaluation pressure	-0.252 (0.271)	-0.204 (0.268)	-0.396 (0.221)	-0.133 (0.233)	0.381 (0.439)	0.323 (0.449)	0.291 (0.356)	0.064 (0.385)
MKT	-0.100* (0.045)	-0.078 (0.044)	-0.030 (0.036)	-0.087* (0.040)	-0.084 (0.044)	-0.073 (0.044)	-0.005 (0.035)	-0.091* (0.038)
Holding an advanced degree	-0.022 (0.102)	-0.116 (0.100)	-0.055 (0.083)	-0.121 (0.089)	-0.055 (0.094)	-0.130 (0.095)	-0.104 (0.077)	-0.119 (0.082)
Total years in teaching	0.018 (0.025)	-0.037 (0.024)	-0.021 (0.020)	-0.022 (0.022)	0.024 (0.024)	-0.028 (0.024)	-0.020 (0.019)	-0.019 (0.021)
Video observation	-0.339** (0.111)	-0.353** (0.109)	-0.064 (0.090)	-0.225* (0.097)	-0.374*** (0.110)	-0.383*** (0.113)	-0.116 (0.090)	-0.215* (0.097)
Teaching a tested- grade	0.202* (0.090)	0.252** (0.088)	0.115 (0.073)	0.080 (0.079)	0.174* (0.089)	0.166 (0.091)	0.078 (0.073)	0.081 (0.078)
The number of students in class	-0.005 (0.010)	-0.014 (0.010)	0.006 (0.008)	0.015 (0.009)	-0.002 (0.010)	-0.004 (0.010)	0.009 (0.008)	0.016 (0.008)
Male	-0.256 (0.175)	-0.307 (0.169)	-0.356* (0.143)	-0.418** (0.155)	-0.166 (0.186)	-0.253 (0.187)	-0.233 (0.151)	-0.406* (0.163)
White	0.300* (0.127)	0.226 (0.123)	0.0553 (0.103)	0.180 (0.111)	0.268* (0.128)	0.210 (0.131)	0.025 (0.105)	0.180 (0.113)
First day of back-to- back observation	0.069* (0.032)	0.066 (0.035)	0.045 (0.031)	0.060 (0.033)	0.067* (0.032)	0.065 (0.035)	0.043 (0.031)	0.059 (0.033)
Type: Individual work	-0.287*** (0.048)	-0.051 (0.052)	-0.153*** (0.046)	-0.172*** (0.049)	-0.284*** (0.048)	-0.049 (0.052)	-0.147** (0.046)	-0.172*** (0.049)
Type: Small group	0.074 (0.044)	0.193*** (0.048)	0.025 (0.042)	0.097* (0.045)	0.071 (0.045)	0.186*** (0.048)	0.022 (0.042)	0.096* (0.045)
First episode	-0.057 (0.065)	-0.095 (0.071)	-0.057 (0.062)	-0.070 (0.066)	-0.056 (0.065)	-0.092 (0.071)	-0.056 (0.062)	-0.070 (0.066)

Table 5 (cont'd)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
Last episode	-0.215*** (0.060)	-0.221*** (0.065)	-0.185** (0.058)	-0.137* (0.062)	-0.218*** (0.060)	-0.226*** (0.065)	-0.188** (0.058)	-0.137* (0.062)
Episode order	-0.059* (0.029)	-0.057 (0.032)	-0.062* (0.028)	-0.073* (0.030)	-0.060* (0.029)	-0.055 (0.032)	-0.062* (0.028)	-0.074* (0.030)
Quadratic term of order	0.004 (0.002)	0.003 (0.003)	0.005* (0.002)	0.005 (0.002)	0.004 (0.002)	0.003 (0.003)	0.005* (0.002)	0.005 (0.003)
ECTs' planning	0.081 (0.060)	0.094 (0.058)	0.041 (0.048)	0.039 (0.052)				
Social norms regarding planning	-0.158 (0.376)	0.155 (0.370)	-0.312 (0.300)	-0.004 (0.319)				
Social norms regarding planning*Evaluation pressure	0.098 (0.165)	0.023 (0.163)	0.180 (0.133)	-0.006 (0.141)				
ECTs' enactment					0.064 (0.054)	0.043 (0.055)	0.065 (0.044)	-0.012 (0.047)
Social norms regarding enactment					0.361 (0.383)	0.342 (0.390)	0.301 (0.310)	0.189 (0.335)
Social norms regarding enactment*Evaluation pressure					-0.188 (0.171)	-0.197 (0.175)	-0.156 (0.138)	-0.081 (0.149)
N	861	853	846	849	861	853	846	849
N of teachers	73	73	73	73	73	73	73	73

Note. Types of episode, video observation, first day of back-to-back observation, first and last episode, teaching tested grade, holding an advanced degree, male, and white are dummy variables. All models included district fixed effects. Coefficients are not standardized because the unit of dependent variable, TRU Math score, is meaningful. * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Discussion

This study is one of the first to examine how teacher evaluation pressure may affect teachers' planning and enactment of ambitious mathematics instruction based on both survey and observation data. Ambitious mathematics instruction has been emphasized by many scholars and policy makers (Lampert et al., 2010), and some researchers have explored how pre-service teachers can develop their skills for enacting such instruction (Kazemi, Franke, & Lampert, 2009). However, in-service teachers may or may not be motivated to enact ambitious mathematics instruction when they face a separate source of pressure, namely, teacher evaluations. ECTs are mostly pre-tenure, which potentially makes them more reactive to teacher evaluation results, since their ratings can determine their job status. In this situation, if teacher evaluation policies motivate teachers to perform in ways that conflict with the enactment of ambitious mathematics instruction, the former would be likely to win out over the latter because teacher evaluations have clear rewards and sanctions.

The current study provides evidence supporting this argument. Teachers who perceive more pressure associated with teacher evaluation tended to move further away from enacting ambitious mathematics instruction. Interestingly, the mathematical aspect of the content covered during the lesson was not affected by evaluation pressure. Instead, ratings for the dimensions of cognitive demand; agency, authority, and identity; and formative assessment were lower when a teacher perceived higher pressure, even after controlling for their pre-measures of the outcome variable. By design, these three dimensions are hard to measure; while raters could quickly rate the first dimension, the mathematics, based on the mathematical task itself, in order to rate the other three aspects, they needed to closely look at subtle interactions between teacher and students. For example, cognitive demand is not just about the task itself. Raters need to analyze how a teacher introduces the task and raises questions for students in order to judge whether the teacher “scaffolds away the challenges” (Schoenfeld et al., 2014). The same experience applies to the formal evaluators for the teachers, usually principals. Although multiple measures of teacher quality are emphasized in new teacher evaluation systems (Grissom & Youngs, 2016),

observation is still the most frequently used part of teacher evaluation (Kraft & Gilmour, 2016). However, in practice, principals may not stay in each class for the entire lesson, due to time constraints.

For non-tenured teachers, some principals have had to spend more than 10 hours per year per teacher for annual evaluation, and they still reported that they had insufficient time for completing the evaluation process (Kersten & Israel, 2005). Principals have also reported that “the new policy exacerbates time requirements” (Ramirez, Clouse, & Davis, 2014, p. 46). Moreover, other components of teacher evaluation, such as student growth, might not be able to capture these three aspects of mathematics instruction easily, because those measures depend on test scores. Taken together, it is arguable that teachers might have made a rational decision to move away from enacting ambitious mathematics instruction and to focus on other aspects of instruction valued by the current teacher evaluation system. This is more evident in those three areas (the three dimensions of the current study besides the mathematics) that evaluators might not be able to easily observe. This is consistent with some teachers’ responses to school accountability policies; with limited time and other resources, teachers strategically spend their time and resources to meet the requirements of *NCLB* (Booher-Jennings, 2005; Jacob, 2005; Reback et al., 2014; White & Rosenbaum, 2008). The current study confirms that the same issue might exist in teacher evaluation settings. Moreover, given that ECTs working in the 2015-16 school year are more likely to experience mathematics methods courses focused on ambitious mathematics instruction in their preparation programs, it is alarming that the effects of these courses are likely diminished when graduates begin working as full-time teachers.

The results about the potential moderating effects of teachers’ MKT and social norms on teachers’ instruction represent how difficult it would be to reconcile the conflict between teacher evaluation and ambitious mathematics instruction. Although teachers’ MKT has been argued to be critical for high-quality mathematics teaching (Hill et al., 2005; Hill et al., 2007), with a high level of teachers’ perceived pressure related to teacher evaluation, the association between their MKT and ambitious mathematics teaching became weaker. Based on the framework that I used, this shows that teachers used their individual resource, MKT, to make their instruction aligned more with teacher evaluation rather than

ambitious mathematics teaching. Since teachers with higher levels of MKT have more resources to use for their instruction, they would be able to change their instruction more dramatically to align it with expectations associated with teacher evaluation. It is also interesting that this moderating effect of MKT was more salient for dimensions 3 and 4, which are dimensions closely related to discussion between students and a teacher. Teachers with high MKT scores tended to have more corrective interactions with students than teachers with lower MKT scores when they faced a high level of evaluation pressure.

On the other hand, this result reveals some shortcomings of studies on teachers' MKT. First, as Hill and colleagues (2007) pointed out, it focuses heavily on the mathematical quality of a lesson, rather than the quality of teaching itself. MKT is essential for teaching, but it is only one element of mathematics instruction. Thus, it cannot guarantee high-quality mathematics teaching, especially when teachers encounter different demands, such as teacher evaluation. In fact, the data for validating MKT surveys was collected before new, more rigorous teacher evaluation policies were enacted. Thus, most studies on MKT do not address how teachers' MKT can be applied in classrooms under current policy dynamics. Second, as a related point, it seems that teachers' MKT is not useful for predicting the quality of a specific lesson. Even in previous research, the correlation between survey-based MKT and lesson MQI (Mathematical Quality of Instruction) was not significant; rather, when a trained rater rated a teacher's MKT based on classroom observation data, it was highly correlated with MQI (Hill et al., 2011). Given that MKT is appealing to researchers partially based on it being a multiple-choice, computerized, short-answer test, this puts some limitations on our ability to understand how it is used in classrooms. In fact, the correlation between MKT and TRU Math scores was not significant in the current study.

In contrast to the expectation that social norms may affect the association between evaluation pressure and teachers' instruction, there were no significant results related to social norms regarding either planning or enactment of ambitious mathematics instruction. This does not mean that social norms regarding mathematics instruction do not have any influence on ECTs' mathematics instruction. Instead, this might mean that there was no buffering effect of social ties on the effects of teacher evaluation policies. This result sharply contrasts with the well-known image of the school organization as a buffer

between external influences and school staff members (Bidwell, 2001). This might be good news for proponents of the current teacher evaluation system, since it shows that teacher evaluations might affect change in instruction, while the results of the current study cast doubt on the direction of this influence.

This point also applies to the results regarding teachers' use of resources in planning. When teachers perceived more pressure associated with teacher evaluation, they were more likely to use resources from outside of the classroom, such as district pacing books, advice from other teachers, and teacher evaluation results and criteria. This again indicates that teacher evaluation policies may shape teachers' instruction. However, connecting this result with those discussed above suggests that teacher evaluation pressure prompts teachers to move further away from enactment of ambitious mathematics instruction by using more resources in planning.

There are some limitations of the current study that can be addressed in future studies. First, this is an observational study including only 84 teachers. It is not possible to draw a firm causal conclusion between teacher evaluation policies and changes in teachers' ambitious mathematics instruction based on this analysis. However, it should be noted that the focus was teachers' perceived pressure associated with teacher evaluation policies, which is almost impossible to examine with experimental studies. Even under the same evaluation systems, each teacher feels differently about their evaluation, which I showed earlier by running an unconditional model; almost 100 percent of the variance in teachers' perceived pressure was at the individual teacher level. Moreover, pre-measures were included in order to reduce potential bias in the analysis.

In addition, it is worthwhile to note Robustness Indices (Frank, 2000; Frank et al., 2008; Frank et al., 2013) associated with the main findings. First, in terms of the association between teachers' perceived pressure related to teacher evaluation and their use of resources in planning (i.e., model 1 in Table 3), the association seems to be strong; to invalidate the inference, 47% of the sample would need to be replaced with cases that do not show a significant association between two variables. The results about the main effects of teacher evaluation pressure on changes in teachers' enactment of ambitious mathematics are relatively weaker. To invalidate the inferences, 15% of the sample for Dimension 2 (Cognitive demand),

6% of the sample for Dimension 3 (Agency, authority, and identity), and 22% of the sample for Dimension 4 (Formative assessment) need to be replaced with cases that show no effects of perceived pressure related to evaluation on teachers' TRU Math scores. In terms of correlation-based indices, an omitted variable would have to be correlated at 0.11 with teachers' perceived pressure related to evaluation and at 0.11 with Dimension 2 scores in Spring after controlling for all covariates including the pre-measure. This correlation is 0.064 for the Dimension 3, and 0.137 for the Dimension 4, respectively. Although these numbers seem to be small, other significant covariates in the model had a similar level of indices. For example, to invalidate the inference that video observations tended to be scored lower than live observations, an omitted variable would have to be correlated only at 0.163 with Dimension 2 scores and at 0.163 with the observation being video. This correlation value for the teachers being male is 0.063 and observations being conducted in the first day is 0.147, respectively. One exception is the dummy variable for the small group activity. The association between an episode being small group and Dimension 2 scores is relatively strong; to invalidate this inference, an omitted variable would have to be correlated at 0.283 with Dimension 2 scores and at 0.283 with the episode being small group.

Another important finding of the current study is about the heterogeneous effects of teachers' perceived pressure related to teacher evaluation on changes in TRU Math scores based on their MKT. According to the Robustness Indices, the association between a term for the interaction between teachers' perceived pressure and MKT and teachers' Dimension 4 scores is quite strong. To invalidate this inference, 40% of the sample would need to be replaced with cases for which the effects of the two variables (i.e., evaluation pressure and MKT) on the outcome are purely additive. According to the correlation framework, an omitted variable would have to be correlated with at 0.212 with Dimension 4 scores and at 0.212 with the interaction term. On the other hand, the association between the interaction term and the Dimension 3 scores was weak; changing only 3% of the sample can invalidate the inference.

Second, the main constructs of the current study, teachers' planning, enactment, perceived evaluation pressure, and social norms, were narrowly defined. All constructs are multi-faceted and complex; it might not be enough to use one or two instruments to measure them. For example, enactment

of ambitious mathematics instruction can unfold in many ways, other than what the TRU Math observation instrument can capture. Similarly, teachers' planning behavior includes many different aspects, other than using resources. Teacher evaluation pressure also can be manifested in different ways; teachers may not even be aware of this pressure. In addition, social norms can be defined in various ways. For example, in this study, TRU Math scores for social network members were not available, so their survey responses were used as a proxy for their planning and enactment of mathematics instruction. Thus, other types of social norms may produce a different result. Taken together, the results from this analysis might apply to a specific area in teachers' planning, enactment, perceived evaluation pressure, and social norms as I defined them, while the variables that I used for this study may not represent all aspects of those constructs.

Based on these limitations, I suggest two directions for future studies. First, qualitative studies that examine the process by which teacher evaluation influences teachers' instruction would be fruitful. What factors affect teachers' perceptions about the pressure associated with teacher evaluation policies? How do such perceptions translate into changes in their instruction? Investigating these questions with in-depth interviews and observations might be one way to capture different aspects of the influence of teacher evaluation policies on teachers.

Second, other factors might affect this association between teacher evaluation pressure and teachers' instruction. To be specific, school leadership can shape the influence of teacher evaluation policies. For example, school administrators who engage in high-quality or learning-centered leadership may weaken or strengthen the influence of teacher evaluation policies, given the critical role of principals in the teacher evaluation process. In addition, districts' other reforms may affect this association; if a district supports ambitious mathematics instruction more actively with a curriculum reform at the district level, teachers may continue to enact such instruction, regardless of teacher evaluation policies. Studying these factors might add new insights for studies on the influence of teacher evaluation policies.

Despite some limitations, this study represents a meaningful effort to analyze the effects of teacher evaluation policies in light of ambitious mathematics instruction. Given the sharp conflict

between what the CCSS and teacher preparation courses stand for, and what teacher evaluation policies expect, ECTs in this study tended to move away from ambitious mathematics instruction as the school year went on. If ambitious mathematics instruction is indeed the most desirable way to teach mathematics, this study provides supporting evidence that current teacher evaluation policies need to be reconsidered.

NOTES

NOTES

¹ A pseudonym is used here.

² Although the larger study administered a survey for principals of ECTs' schools, the response rate was lower than 30 percent, so I did not include principals' data in the current study.

³ The last item was not included in the dissertation proposal. Without this item, the Cronbach alpha is too low ($\alpha=0.484$) and the correlation among items is too low to compose a factor (Eigen value=0.581 for the first factor). Appendix B reports the results using only three items, but the results remain similar.

APPENDICES

Appendix A

TRU Math Rubric

Table 6. TRU Math Summary Rubric

The Mathematics	Cognitive Demand	Agency, Authority, and Identity	Formative Assessment
<i>How accurate, mathematically rich (with regard to content and practices), focused, coherent, and well justified is the lesson content?</i>	<i>To what extent are students supported in grappling with and making sense of mathematical concepts?</i>	<i>To what extent are students the source of ideas and discussion of them? How are student contributions framed?</i>	<i>To what extent is students' mathematical thinking surfaced; to what extent do subsequent interactions build on student ideas when potentially valuable or address misunderstandings when they arise?</i>
Classroom activities are unfocused or skills-oriented, lacking opportunities for engagement with key grade level content (as specified in the Common Core Standards)	Classroom activities are structured so that students mostly apply memorized procedures and/or work routine exercises.	Students' speech turns may be short (one sentence or less), and constrained by what the teacher or materials direct them to do.	The teacher does not actively surface or pursue student thinking. Teacher and student actions are limited to corrective feedback or encouragement.
Activities are at grade level but are primarily skills-oriented, with little overt support for making connections, for mathematical coherence (see glossary), or for engaging in mathematical practices.	Classroom activities offer possibilities of conceptual richness or problem solving challenge, but interactions tend to "scaffold away" the challenges, removing opportunities for productive struggle.	Students have a chance to explain some of their thinking, but the teacher is the primary driver of conversations and arbiter of correctness. The teacher may ascribe ownership for students' ideas, but students do not build upon each other's ideas.	The teacher surfaces student reasoning, but specific students' ideas are not built on (when potentially valuable) or used to address challenges (when problematic).
Classroom activities include meaningful connections and provide opportunities for building a coherent view of mathematics, including content and practices.	The teacher's hints/scaffolds or student-student interactions support students in productive struggle in building understandings and engaging in mathematical practices.	Students explain their ideas and reasoning AND students respond to and build on each other's ideas.	The teacher surfaces and explores student reasoning, or student comments respond to those ideas, by building on productive beginnings or addressing emerging misunderstandings.

Note. This is a summary rubric across all activity types, and the SAMI project uses activity type-specific rubrics for coding. Source: Schoenfeld, A. H., Floden, R. E., & the Algebra Teaching Study and Mathematics Assessment Project. (2014). *The TRU Math Scoring Rubric*. Berkeley, CA & E. Lansing, MI: Graduate School of Education, University of California, Berkeley & College of Education, Michigan State University. Retrieved from <http://ats.berkeley.edu/tools.html>.

Appendix B

Results Using Three Evaluation Pressure Items

Table 7. Potential Effects of Teachers' Perceived Pressure Related to Teacher Evaluation on Their Use of Resources in Planning (Using three items)

	Model (1)	Model (2)	Model (3)	Model (4)
Use of resource in fall 2015	0.519*** (0.088)	0.519*** (0.088)	0.533** (0.108)	0.493*** (0.083)
Evaluation pressure	0.365** (0.115)	0.363** (0.085)	0.768* (0.317)	0.320 (0.757)
MKT	-0.221*** (0.021)	-0.200 (0.310)	-0.179* (0.036)	-0.192* (0.042)
MKT*Evaluation pressure		-0.022 (0.137)		
Holding an advanced degree	0.100 (0.079)	0.101 (0.077)	0.118 (0.114)	0.166 (0.156)
Total years in teaching	-0.123 (0.024)	-0.123 (0.027)	-0.192* (0.027)	-0.165* (0.020)
Male	0.022 (0.161)	0.023 (0.148)	0.026 (0.222)	-0.001 (0.231)
White	-0.070 (0.135)	-0.070 (0.135)	-0.119 (0.135)	-0.153 (0.164)
Teaching a tested-grade	-0.058 (0.122)	-0.059 (0.126)	-0.008 (0.054)	0.030 (0.090)
The number of students in class	-0.025 (0.018)	-0.024 (0.017)	-0.114 (0.009)	-0.122 (0.014)
Social norm regarding planning			0.437 (0.489)	
Social norm regarding enactment				0.006 (0.625)
Social norm regarding planning*Evaluation pressure			-0.563 (0.187)	
Social norm regarding enactment*				0.032 (0.278)
Evaluation pressure				
Enactment in fall 2015				0.050 (0.067)
R-squared	0.615	0.615	0.639	
N	79	79	74	

Note. All models included district fixed effects. All coefficients are standardized and clustered robust standard errors at district level are in parentheses. Enactment in fall 2015 is included in model 4 to control for potential selection effect by ECTs' enactment level. * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 8. Potential Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation on Teachers' Enactment of Mathematics
Instruction: Main Effects and Heterogeneous Effects Based on Teachers' MKT (Using three items)

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Formative assessment	Model (5) Dimension1: The mathematics	Model (6) Dimension2: Cognitive demand	Model (7) Dimension3: Agency, authority, and identity	Model (8) Dimension4: Formative assessment
Mean score in fall 2015	0.137 (0.100)	0.085 (0.088)	0.135 (0.088)	0.149 (0.084)	0.137 (0.095)	0.086 (0.087)	0.151 (0.087)	0.170* (0.082)
Evaluation pressure	-0.019 (0.068)	-0.112 (0.064)	-0.084 (0.051)	-0.114* (0.053)	-0.040 (0.073)	-0.137* (0.069)	-0.112* (0.054)	-0.162** (0.055)
MKT	-0.070 (0.041)	-0.048 (0.040)	0.003 (0.031)	-0.068* (0.033)	0.075 (0.194)	0.129 (0.186)	0.219 (0.149)	0.296 (0.152)
Holding an advanced degree	-0.161 (0.091)	-0.132 (0.087)	-0.114 (0.070)	-0.116 (0.073)	-0.162 (0.0906)	-0.134 (0.087)	-0.115 (0.069)	-0.119 (0.071)
Total years in teaching	0.019 (0.025)	-0.016 (0.024)	-0.013 (0.019)	-0.019 (0.020)	0.016 (0.025)	-0.019 (0.024)	-0.017 (0.019)	-0.026 (0.020)
Video observation	-0.311** (0.106)	-0.270** (0.103)	-0.007 (0.082)	-0.149 (0.087)	-0.322** (0.107)	-0.283** (0.103)	-0.022 (0.082)	-0.175* (0.085)
Teaching a tested grade	0.035 (0.084)	0.061 (0.081)	-0.006 (0.064)	0.010 (0.067)	0.029 (0.084)	0.053 (0.081)	-0.015 (0.064)	-0.005 (0.066)
The number of students in class	-0.001 (0.010)	-0.0001 (0.010)	0.017* (0.008)	0.018* (0.008)	-0.001 (0.010)	0.0004 (0.010)	0.017* (0.008)	0.019* (0.008)
Male	-0.333* (0.147)	-0.275* (0.140)	-0.281* (0.112)	-0.356** (0.117)	-0.321* (0.148)	-0.259 (0.140)	-0.260* (0.111)	-0.320** (0.115)
White	0.298* (0.131)	0.192 (0.126)	0.0351 (0.101)	0.143 (0.105)	0.309* (0.131)	0.205 (0.126)	0.051 (0.101)	0.172 (0.103)
First day of back-to-back observation	0.079** (0.029)	0.084** (0.032)	0.055 (0.029)	0.067* (0.030)	0.079** (0.029)	0.084** (0.032)	0.055 (0.029)	0.067* (0.030)
Type: Individual work	-0.245*** (0.044)	0.015 (0.048)	-0.153*** (0.043)	-0.158*** (0.046)	-0.245*** (0.044)	0.015 (0.048)	-0.153*** (0.043)	-0.160*** (0.046)
Type: Small group	0.064 (0.040)	0.194*** (0.044)	0.0422 (0.039)	0.116** (0.042)	0.063 (0.040)	0.194*** (0.044)	0.0407 (0.039)	0.113** (0.041)

Table 8 (cont'd)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
First episode	-0.037 (0.059)	-0.084 (0.065)	-0.110 (0.058)	-0.116 (0.062)	-0.037 (0.059)	-0.083 (0.065)	-0.109 (0.058)	-0.115 (0.062)
Last episode	-0.205*** (0.054)	-0.185** (0.059)	-0.178*** (0.054)	-0.134* (0.057)	-0.207*** (0.054)	-0.187** (0.059)	-0.182*** (0.054)	-0.141* (0.057)
Episode order	-0.047 (0.027)	-0.050 (0.030)	-0.073** (0.027)	-0.077** (0.028)	-0.047 (0.027)	-0.050 (0.030)	-0.073** (0.027)	-0.077** (0.028)
Quadratic term of order	0.003 (0.002)	0.002 (0.003)	0.005* (0.002)	0.005 (0.002)	0.003 (0.002)	0.002 (0.003)	0.005* (0.002)	0.005* (0.002)
MKT*Evaluation pressure					-0.069 (0.090)	-0.084 (0.087)	-0.103 (0.069)	-0.174* (0.071)
N	986	978	972	976	986	978	972	976
N of teachers	84	84	84	84	84	84	84	84

Note. Types of episode, video observation, first day of back-to-back observation, first and last episode, teaching tested grade, holding an advanced degree, male, and white are dummy variables. All models included district fixed effects. Coefficients are not standardized because the unit of dependent variable, TRU Math score, is meaningful. In models 5 through 8, teachers' MKT and evaluation pressure are grand-mean centered.

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 9. Potential Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation on Teachers' Enactment of Mathematics
Instruction: Heterogeneous effects Based on Social Norms (Using three items)

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Formative assessment	Model (5) Dimension1: The mathematics	Model (6) Dimension2: Cognitive demand	Model (7) Dimension3: Agency, authority, and identity	Model (8) Dimension4: Formative assessment
Mean score in fall 2015	0.199* (0.094)	0.095 (0.088)	0.256** (0.0995)	0.257** (0.010)	0.176* (0.089)	0.080 (0.090)	0.254* (0.099)	0.260** (0.100)
Evaluation pressure	-0.172 (0.269)	-0.087 (0.265)	-0.360 (0.220)	0.032 (0.232)	0.695* (0.349)	0.543 (0.362)	0.304 (0.288)	0.127 (0.315)
MKT	-0.097* (0.045)	-0.075 (0.044)	-0.027 (0.036)	-0.080* (0.039)	-0.082 (0.043)	-0.071 (0.044)	-0.003 (0.035)	-0.090* (0.038)
Holding an advanced degree	-0.026 (0.099)	-0.116 (0.098)	-0.062 (0.081)	-0.128 (0.087)	-0.052 (0.092)	-0.124 (0.096)	-0.101 (0.077)	-0.112 (0.083)
Total years in teaching	0.0175 (0.025)	-0.040 (0.024)	-0.023 (0.020)	-0.027 (0.022)	0.028 (0.024)	-0.025 (0.024)	-0.019 (0.020)	-0.019 (0.022)
Video observation	-0.330** (0.110)	-0.334** (0.109)	-0.061 (0.089)	-0.216* (0.097)	-0.364*** (0.108)	-0.366** (0.112)	-0.109 (0.090)	-0.201* (0.098)
Teaching a tested grade	0.182* (0.090)	0.220* (0.089)	0.101 (0.074)	0.052 (0.079)	0.168* (0.084)	0.155 (0.088)	0.0774 (0.071)	0.0691 (0.076)
The number of students in class	-0.005 (0.010)	-0.014 (0.010)	0.005 (0.008)	0.015 (0.009)	-0.001 (0.010)	-0.005 (0.010)	0.009 (0.008)	0.015 (0.009)
Male	-0.234 (0.178)	-0.279 (0.173)	-0.345* (0.145)	-0.382* (0.157)	-0.124 (0.175)	-0.225 (0.180)	-0.237 (0.145)	-0.395* (0.159)
White	0.320* (0.130)	0.239 (0.128)	0.0733 (0.106)	0.176 (0.114)	0.315* (0.129)	0.244 (0.134)	0.0389 (0.108)	0.191 (0.117)
First day of back-to- back observation	0.069* (0.032)	0.067 (0.035)	0.045 (0.031)	0.061 (0.033)	0.067* (0.032)	0.066 (0.035)	0.044 (0.031)	0.060 (0.033)
Type: Individual work	-0.287*** (0.048)	-0.050 (0.052)	-0.154*** (0.046)	-0.172*** (0.049)	-0.281*** (0.048)	-0.047 (0.052)	-0.146** (0.046)	-0.170*** (0.049)
Type: Small group	0.076 (0.044)	0.198*** (0.048)	0.028 (0.042)	0.106* (0.045)	0.066 (0.045)	0.186*** (0.048)	0.023 (0.042)	0.100* (0.045)

Table 9 (cont'd)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
First episode	-0.057 (0.065)	-0.095 (0.071)	-0.058 (0.062)	-0.070 (0.066)	-0.056 (0.065)	-0.092 (0.071)	-0.056 (0.062)	-0.071 (0.066)
Last episode	-0.214*** (0.060)	-0.218*** (0.065)	-0.184** (0.058)	-0.134* (0.062)	-0.219*** (0.060)	-0.226*** (0.065)	-0.187** (0.058)	-0.135* (0.062)
Episode order	-0.060* (0.029)	-0.057 (0.032)	-0.062* (0.028)	-0.075* (0.030)	-0.059* (0.029)	-0.055 (0.032)	-0.062* (0.028)	-0.074* (0.030)
Quadratic term of order	0.004 (0.002)	0.003 (0.003)	0.005* (0.002)	0.005 (0.002)	0.004 (0.002)	0.003 (0.003)	0.005* (0.002)	0.005 (0.002)
ECTs' planning	0.091 (0.060)	0.111 (0.059)	0.051 (0.048)	0.058 (0.052)				
Social norms regarding planning	-0.101 (0.354)	0.245 (0.347)	-0.285 (0.284)	0.167 (0.302)				
Social norms regarding planning*Evaluation pressure	0.0727 (0.159)	-0.0219 (0.157)	0.170 (0.129)	-0.0903 (0.137)				
ECTs' enactment					0.056 (0.053)	0.045 (0.055)	0.069 (0.044)	-0.005 (0.048)
Social norms regarding enactment					0.609* (0.310)	0.506 (0.322)	0.299 (0.256)	0.222 (0.280)
Social norms regarding enactment*Evaluation pressure					-0.307* (0.141)	-0.277 (0.146)	-0.158 (0.116)	-0.0973 (0.127)
N	861	853	846	849	861	853	846	849
N of teachers	73	73	73	73	73	73	73	73

Note. Types of episode, video observation, first day of back-to-back observation, first and last episode, teaching tested grade, holding an advanced degree, male, and white are dummy variables. All models included district fixed effects. Coefficients are not standardized because the unit of dependent variable, TRU Math score, is meaningful. * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

REFERENCES

REFERENCES

- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389-407.
- Bidwell, C. E. (2001). Analyzing schools as organizations: Long-term permanence and short-term change. *Sociology of Education*, 74, 100-114.
- Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, 42(2), 231-268.
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Center for Educational Leadership. (n.d.). 5 dimensions of teaching learning. Seattle, VA: Authors.
- Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, 23(2), 145-170.
- Coggshall, J. G., Rasmussen, C., Colton, A., Milton, J., & Jacques, C. (2012). *Generating teaching effectiveness: the role of job-embedded professional learning in teacher evaluation. Research & Policy Brief*. National Comprehensive Center for Teacher Quality. Retrieved from <http://eric.ed.gov/?id=ED532776>
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25(2), 119-142.
- Colby, S. A., Bradshaw, L. K., & Joyner, R. L. (2002). Teacher evaluation: A review of the literature. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724-750.
- Danielson, C. (1996). Enhancing professional practice: A framework for teaching. Alexandria, VA: Association of Supervision and Curriculum Development.
- Davis, D. R., Ellett, C. D., & Annunziata, J. (2002). Teacher evaluation, leadership and learning organizations. *Journal of Personnel Evaluation in Education*, 16(4), 287-301.
- Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30(3), 418-446.
- Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching & Teacher Education*, 36, 1-11.
- Donaldson, M. (2012). *Teachers' perspectives on teacher evaluation reform*. Washington, DC: Center for American Progress.

- Donaldson, M. L., & Papay, J. P. (2012). *Reforming teacher evaluation: One district's story*. Washington, D.C: Center for American Progress.
- Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Palo Alto, CA: Stanford University Press.
- Firestone, W. A., Blitz, C. L., Gitomer, D.H., Kirova, D., Shcherbakov, A., & Nordon, T. L. (2013) *New Jersey teacher evaluation, RU-GSE external assessment, year 1 report*, NJ: Rutgers University-Graduate School of Education.
- Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research*, 29(2), 147-194.
- Frank, K. A., Kim, C., & Belman, D. (2010). Utility theory, social networks, and teacher decision making. In A. J. Daly (Ed.), *Social network theory and educational change* (pp.223–242). Cambridge, MA: Harvard University Press.
- Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, 35(4), 437-460.
- Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? *Education Evaluation and Policy Analysis*, 30(1), 3–30.
- Frank, K. A., Zhao, Y., Penuel, W. R., Ellefson, N., & Porter, S. (2011). Focus, fiddle and friends: Experiences that transform knowledge for the implementation of innovations. *Sociology of Education*, 84(2), 137-156.
- Geijsel, F., Sleegers, P., Berg, R. V, & Kelchtermans, G. (2001). Conditions fostering the implementation of large-scale innovation programs in schools: Teachers' perspectives. *Educational Administration Quarterly*, 37(1), 130–166.
- Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, 44(2), 87–95.
- Grissom, J. A., & Youngs, P. (Eds.). (2016). *Improving teacher evaluation systems: Making the most of multiple measures*. New York: Teachers College Press.
- Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational Administration, Policy, and Reform: Research and Measurement*, 3, 153–188.
- Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., & Robyn, A. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: Rand Corporation.
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
- Harris, D. N., & Herrington, C. D. (2015). The use of teacher value-added measures in schools: New

- evidence, unanswered questions, and future prospects. *Educational Researcher*, 44(2), 71–76.
- Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the ecological assumption: The relationship of measure scores to classroom teaching and student learning. *Measurement*, 5(2-3), 107-118.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794-831.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371-406.
- Hill, H.C., Schilling, S.G., & Ball, D.L. (2004) Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal* 105(1), 11-30.
- Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal*, 52(2), 208–242.
- Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, 89(5), 761–796.
- Jacob, B. A., & Levitt, S. D., (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 843–77.
- Jiang, J. Y., Spote, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform Chicago's REACH students. *Educational Researcher*, 44(2), 105–116.
- Kazemi, E., Franke, M., & Lampert, M. (2009, July). Developing pedagogies in teacher education to support novice teachers' ability to enact ambitious instruction. In *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 12-30). Adelaide, SA: MERGA.
- Kersten, T. A., & Israel, M. S. (2005). Teacher evaluation: Principals' insights and suggestions for improvement. *Planning and Changing*, 36(1/2), 47-67.
- Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education*, 16(4), 241–268.
- Lampert, M., Beasley, H., Ghousseini, H., Kazemi, E., & Franke, M. (2010). Using designed instructional activities to enable novices to manage ambitious mathematics teaching. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 129-141). New York: Springer.
- McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, 9(2), 171–178.
- Milanowski, A. T., & Herbert, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, 15(3), 193–212.
- Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation.

Harvard Educational Review, 82(1), 123–141

Penuel, W. R., Frank, K. A., Sun, M., Kim, C., & Singleton, C. (2013). The organization as a filter of institutional diffusion. *Teachers College Record*, 115(1), 306-339.

Ramirez, A., Clouse, W., & Davis, K. W. (2014). Teacher evaluation in Colorado: How policy frustrates practice. *Management in Education*, 28(2), 44-51.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Reback, R. (2008) Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5), 1394–1415.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, 6(3), 207–241.

Riordan, J., Lacireno-Paquet, N., Shakman, K., Bocala, C., & Chang, Q. (2015). *Redesigning teacher evaluation: Lessons from a pilot implementation* (REL 2015–030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from <http://ies.ed.gov/ncee/edlabs>

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

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy*, 5(2), 251–281.

Salloum, S.J., Bieda, K.N., Sweeny, S.P., Torphy, K.T., Hu, S., & Lane, J. (2016). Capturing early career teachers' enactment of ambitious mathematics practice at scale. Paper presented at the annual meeting of the American Educational Research Association, Washington D.C.

Schneider, A., & Ingram, H. (1990). Behavioral assumptions of policy tools. *The Journal of Politics*, 52(2), 510–529.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM*, 45(4), 607-621.

Schoenfeld, A. H., Floden, R. E., & the Algebra Teaching Study and Mathematics Assessment Project. (2014). *The TRU Math scoring rubric*. Berkeley, CA & E. Lansing, MI: Graduate School of Education, University of California, Berkeley & College of Education, Michigan State University. Retrieved from <http://ats.berkeley.edu/tools.html>.

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79-100). Santa Monica: RAND Corporation.

Stronge, J. H. (1995). Balancing individual and institutional goals in educational personnel evaluation: A conceptual framework. *Studies in Educational Evaluation*, 21(2), 131–151.

- Sun, M., Frank, K. A., Penuel, W. R., & Kim, C. M. (2013). How external institutions penetrate schools through formal and informal leaders. *Educational Administration Quarterly*, 49(4), 610-644.
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628–3651.
- Thames, M. H., & Ball, D. L. (2010). What math knowledge does teaching require? *Teaching Children Mathematics*, 17(4), 220-229.
- Tuytens, M., & Devos, G. (2009). Teachers' perception of the new teacher evaluation policy: A validity study of the policy characteristics scale. *Teaching and Teacher Education*, 25(6), 924–930.
- Tuytens, M., & Devos, G. (2010). The influence of school leadership on teachers' perception of teacher evaluation policy. *Educational Studies*, 36(5), 521–536.
- Waters Public Schools (2012). *Professional evaluation system*. Waters, IN: Waters Public Schools.
- White, K. & Rosenbaum, J. (2008). Inside the black box of accountability: How high-stakes accountability alters school culture and the classification and treatment of students and teachers. In A. R. Sadovnik, J. A. O'Day, G. W. Bohrnstedt, & K. M. Borman (Eds.), *No Child Left Behind and the reduction of the achievement gap: Sociological perspectives on federal education policy* (pp. 97–116). New York: Routledge.
- Youngs, P., Frank, K.A., & Pogodzinski, B. (2012). The role of mentors and colleagues in beginning elementary and middle school teachers' language arts instruction. In S. Kelly (Ed.), *Understanding teacher effects* (pp.161-181). New York: Teachers College Press.
- Youngs, P., & Haslam, M. B. (2012). A review of research on emerging teacher evaluation systems. Washington, DC: Policy Studies Associates.

Essay 2: Teacher Evaluation Policies in a Loosely Coupled System: Their Implementation and Effects in Michigan School Districts

Over the last decades, policymakers in the U.S. have exerted unprecedented pressure to shift the educational system from a loosely coupled to a tightly coupled system (Lowe Boyd & Crowson, 2002; Fusarelli, 2002; Meyer, 2002). Despite the unresolved debate about whether loose coupling is a serious problem that needs to be addressed or if it is an unchangeable reality of school organizations (Orton & Weick, 1990; Shen, Gao, & Xia, 2016), such environmental pressure promotes tight couplings of the education system in the U.S. (Fusarelli, 2002). The *No Child Left Behind Act* (NCLB) is a representative manifestation of such a movement with its focus on the “technical core of schooling” (Meyer & Rowan, 1977), through standard-based curricula, testing, and scrutiny of testing outcomes followed by rewards and sanctions. In turn, teacher evaluation policies have also drawn significant attention, as it has been shown that teaching quality is one of the most important factors influencing student learning (Aaronson, Barrow, & Sander, 2007; Koedel & Betts, 2007; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain 2005; Rockoff, 2004; Sanders, Wright, & Horn, 1997). In response to the federal emphasis on teacher evaluation in Race to the Top and ESEA Title 1 waivers, many states have passed legislation to establish new, more rigorous teacher evaluation systems with multiple measures of teaching quality and an emphasis on students’ achievement scores (Ballou & Springer, 2015; Harris, Ingle, & Rutledge, 2014; Herlihy et al., 2014; Pogodzinski, Umpstead, & Witt, 2015; Steinberg & Sartain, 2015). From loose coupling theory, such emphasis on teacher evaluation policies is clearly a movement toward a tightly coupled system. As defined by Weick (1982), inspection combined with feedback for personnel performance is one of the main features of tightly coupled systems.

However, the pendulum swung back to loose coupling recently; the *Every Student Succeeds Act* (ESSA) in 2015 granted states more flexibility with regard to school accountability, and states are now left to make decisions about the future of their teacher evaluation policies (National Education Association, 2015). In March 2017, the U.S. Secretary of Education Betsy DeVos emphasized maintaining high levels of local control over schooling, releasing an updated ESSA consolidated state

plan template (U.S. Department of Education, 2017). Although some states will keep playing the same dominant role in teacher evaluation policies that the federal government used to play, other states may give local school districts more flexibility, allowing them to decide on specific aspects of teacher evaluation. In fact, allowing states to make their own decisions about their roles in evaluation procedures indicates a departure from the previous movement toward tightly coupled systems. That is, although teacher evaluation policies themselves represent the idea of tight coupling, they are often implemented in loosely coupled ways. Under this circumstance, the role of teacher evaluation regarding tightening/loosening the system depends on the implementation of the policies in different jurisdictions. As Ingersoll (1993) argued, “(n)ecessary now is more systematic and detailed investigation into the questions of to what degree, in regard to which organizational tasks, under what conditions, in which organizations and with what consequences, which forms of tight and loose coupling hold” (p. 42).

Given the current policy movement towards local control over teacher evaluation policies, school districts in Michigan provide an optimal environment for examining the implementation of such policies. Michigan has a long history of local control and school districts in Michigan have developed their own instructional policies for a long time (Spillane, 1996). In order to be competitive for Race to the Top funds and qualify for an ESEA waiver, however, Michigan enacted a new state-level teacher evaluation policy in 2010. In addition, the state required each district to enact teacher evaluation policies and to report teachers’ ratings beginning in the 2011-12 school year (Keesler & Howe, 2016). In contrast to other states, which already had detailed state-level laws related to teacher evaluation policies, the Michigan Department of Education (MDE) provided districts with a high level of autonomy in choosing classroom observation tools and student growth measures, training evaluators and teachers, and selecting other components of teacher evaluation, such as student surveys (Michigan Department of Education, 2014). In addition, recently, the Michigan Senate approved legislation that puts more emphasis on local control over teacher evaluation policies and lowers the importance of student growth in teachers’ summative evaluations (Oosting, 2015). In short, teacher evaluation policies in Michigan school districts unfolded in a very unique setting in light of the framework of loose coupling: 1) teacher evaluation

policies were enacted in a historically loosely coupled system; 2) such policies were one of the main instruments for tight coupling; but 3) without clear guidelines, the implementation of the policies seems to be loosely coupled.

In this regard, my first research question focuses on the degree of variation in the ways in which Michigan school districts implemented teacher evaluation policies. In other words, are Michigan school districts still loosely coupled with the state government even as they are implementing policies that aim for tight coupling? To be specific, I focus on two aspects of the policies: the timing of the implementation and whether districts used teacher evaluation ratings to make decisions about teacher dismissal. The timing of the implementation is related to whether the system is tightly or loosely coupled; in loosely coupled systems, events or activities that are happening in one part of the system are not necessarily happening in other parts (Firestone, 1985; Gamoran & Dreeben, 1986). That is, the existence of early adopters, late adopters, and non-compliers indicates the looseness of the coupling in a given system. In terms of the use of teacher evaluation ratings, Hallinger, Heck, and Murphy (2014) pointed out three paths in a theory of action underlying teacher evaluation policies: 1) filter out poor performers, 2) provide feedback and support, and 3) create results-oriented school cultures. Dismissal of teachers based on teacher evaluation ratings is related to the first aspect. It is an intended mechanism for achieving the goal of the policies. However, compared with the other two aspects, filtering out poor performers might be the most challenging part to implement, in that it involves legal issues, relationships with teachers' unions, and resistance from teachers. Despite these challenges, this part would be implemented if districts are tightly coupled with the state government.

The next research question is which factors might have affected such decisions within school districts. That is, which factors make a given district more responsive to the state-level decision? Informed by interview data from 11 districts administrators and two Michigan Education Association representatives, I examined fiscal resources, leadership, student achievement, and demographics as potential factors that affect districts' decision making regarding the implementation of teacher evaluation policies.

The second part of this essay examines the effects of teacher evaluation policies on student achievement. Answering this question is important for two reasons. First, theoretically, it can answer the question of whether a policy that addresses a technical core of teaching can improve student achievement in a historically loosely coupled system without a clearly unified implementation process. Second, practically, there have been few research studies on the effects of teacher evaluation policies at the district level. Although some studies based on teachers' individual data have shown some positive effects of teacher evaluation (Steinberg & Sartain, 2015; Taylor & Tyler, 2012), it is possible that the cost of implementing the policies outstrips their effects at the district level. Moreover, most of the existing literature has focused on one or two components of teacher evaluation policies, such as classroom observations or teacher value-added measures (e.g., Taylor & Tyler, 2012), and/or relied on the piloting of policies or an experimental situation, rather than studying actual policy implementation (e.g., Measures of Effective Teaching project; Steinberg & Sartain, 2015). However, in practice, various components of evaluation (i.e., observation, students' growth/value-added, teachers' self-appraisal, and students' survey) are jointly implemented, and districts need to implement the policies within the limits of available resources as opposed to being supported by external research funding. More importantly, teacher evaluation results can be linked to teachers' future job status in practice. In other words, despite long-standing debate about this topic, the effects of the policies on student achievement at the district level are still unclear.

Accordingly, I focus on the causal relationship between the implementation of teacher evaluation policies and student achievement in Michigan school districts. Combining survey data from 101 districts and MEAP (Michigan Educational Assessment Program) test scores at the district level and applying a quasi-experimental approach, an Interrupted Time Series (ITS), I examined the impact of the implementation of teacher evaluation on student achievement in mathematics and reading. It should be noted that this essay focuses on the effects of fully implementing policies as defined by the state law, as compared to partial implementation of the policies. The requirement of the 2011 legislation (HB 4627) includes three main components: 1) districts need to evaluate all teachers and administrators, 2) annually,

and 3) use evaluation ratings for making different decisions (2011). The first and second components are different in that some districts required tenured teachers to be evaluated once in three years or so. In this case, the district did not fully enact the policies. I define the timing of enactment in each district as occurring when a district started all three components of teacher evaluation. For example, if a district evaluated all teachers annually for the first time in the 2010-11 school year, but started to use teacher evaluation results in the 2012-13 school year, enactment for this district occurred in the 2012-13 school year. In other words, the results from my analysis can be treated as a lower bound of the effects of teacher evaluation at the district level. In addition to the effects on the achievement levels of all students, I also analyzed the effects of the policies on different groups of students separately, such as female/male, economically disadvantaged/non-disadvantaged, White/Black/Hispanic, and elementary/middle school students. As Steinberg and Sartain (2015) showed, teacher evaluation can have different effects across different groups of students.

This study is one of the first to examine the implementation of teacher evaluation policies and their effects on student achievement at the district level in a state context where districts have a high level of local control. In order to more fully understand how the policies were implemented, this study utilizes unique data that have rarely been used in the previous literature: district-level administrators' reports of their implementation of teacher evaluation policies. Although MDE requires districts to report teacher evaluation ratings (i.e., aggregated ratings at the district level), which evaluation tools they used, and which components of teacher evaluation were included (e.g., professionalism, professional development, classroom management, etc.) in every year's data, there was no information available regarding when each district enacted the teacher evaluation policies as the law demanded. That is, the implementation process in each district has rarely been monitored by state government, which also indicates the looseness of the system. The rest of this essay is organized as follows: the following sections introduce the theoretical framework used in this study, loose coupling, and review relevant literature about teacher evaluation and district decision making. The next section describes Michigan's teacher evaluation policies followed by methods and results sections. The last section presents discussion and implications.

Theoretical Framework

The ideas of “loose coupling” (Weick, 1976) and “structural looseness” (Bidwell, 1965) have been widely used by researchers in examining school organizations (Murphy & Hallinger, 1988; Orton & Weick, 1990). In particular, this framework has drawn significant attention due to its utility in describing how schools actually operate, which other classic frameworks, such as bureaucracy or organizational rationality, often failed to do (Firestone, 1985; Fusarelli, 2002; Ingersoll, 1993). Weick’s (1976) famous analogy of a soccer field represents the image of loosely coupled systems vividly: “there are several goals scattered haphazardly around the circular field...and the game is played as if it makes sense” (p. 1). Applying this image to school organizations is straightforward; people have their own purposes and goals in participating in schooling and they make sense of their tasks and the situation as they work. However, they are still on the same field, conducting similar sorts of activities. Although Weick (1976) applied this image to schools that consist of a principal, teachers, students, and parents, it is also possible that loose coupling occurs between hierarchical levels (Orton & Weick, 1990), such as districts and state-level agents as in this study.

Weick (1976) defined loose coupling as the notion that “coupled events are responsive, but (that) each event also preserves its own identity and some evidence of its physical or logical separateness” (p. 3). Building on this idea, Orton and Weick (1990) argued that researchers should use a dialectical interpretation of loose coupling rather than a unidimensional one. Dialectical interpretation of loose coupling puts emphasis on the both distinctiveness and responsiveness of the system: “if there is neither responsiveness nor distinctness, the system is not really a system, and it can be defined as a non-coupled system. If there is responsiveness without distinctiveness, the system is tightly coupled. If there is distinctiveness without responsiveness, the system is decoupled. If there is both distinctiveness and responsiveness, the system is loosely coupled” (Orton & Weick, 1990, p. 205). In contrast, a unidimensional interpretation defines loosely coupled systems as those that have independent components without responsiveness. This dialectical definition is aligned with the argument that the looseness of coupling can vary across different aspects of the same organization; for example, teacher certification and

teachers' pay are tightly coupled, while the technical core of schooling is loosely coupled (Elmore, 2000; Firestone, 1985; Meyer & Rowan, 1977; Weick, 1982).

To be sure, some aspects of the technical core of schooling also seem to become tightly coupled in response to external pressures to tight coupling along with the strong influence of state and federal level agents (Fusarelli, 2002). This discussion indicates that it might not be useful to debate whether a current school system, as a whole, is loosely coupled or tightly coupled. Instead, questioning which aspects of the school system are loosely coupled under which conditions and how the current system can support student learning may be more important (Ingersoll, 1993). Accordingly, I focus on 1) school districts and state-level agents, in terms of the levels of the system; and 2) the implementation of teacher evaluation policies, in terms of a specific aspect of schooling, rather than discussing whether entire school systems are loosely coupled or tightly coupled.

Buffering has been regarded as one of the biggest advantages and, at the same time, disadvantages of loosely coupled systems. According to Weick, “(t)ightly coupled systems overreact to small disturbances (everyone is affected by everything), and loosely coupled systems underreact to large disturbances (no one is affected by anything)” (1982, p. 674). Thus, loose coupling contributes to maintaining the current form of an organization, while it curbs new changes and efforts to improve the organization (Mayer & Rowan, 1977). Proponents of loose coupling assume that this is an unchangeable reality of school organizations that needs to be accepted. In order to make any changes in school organizations, the intervention should be aligned with the fact that they are loosely coupled (Elmore, 2000; Goldspink, 2007; Meyer, 2002). In contrast, opponents argue that we can and should solve this issue of loose coupling by transforming school organizations into tightly coupled systems, in order to improve their effectiveness (Fusarelli, 2002; Lutz, 1982).

Teacher evaluation policies, as a part of standards-based reform, can be considered in two different ways using this lens. For proponents of loose coupling, the policies might not produce any effects on student learning as they go against the loose coupling of current school organizations, and more seriously, they could reduce the support for and legitimacy of the entire organizations, as they “hit at a

critical weakness of the existing institutional structure” (Elmore, 2000, p. 9). On the other hand, opponents would argue that the policies contribute to tightening the system, which is conducive to student learning as tight inspections of performance become possible. Moreover, as the policies move school organizations toward more tightly coupled systems, the whole educational system would become more manageable by authorities at the state or federal levels, and future initiatives would be easier to enact.

In this study, I add another layer to this discussion; in teacher evaluation settings, the core of the policy idea and the implementation of it are conceptualized differently by loose coupling theory. As noted above, the policies themselves have been initiated as part of the movement toward tightly coupled systems, while the policies in Michigan lack detailed rules for implementation and open possibilities for loose coupling. My question here is, specifically in teacher evaluation settings, whether districts are still loosely coupled with the state-level government, and, if so, can we better ascertain what factors might affect districts’ implementation of the policies, and how they affect student learning.

Literature Review

This essay examines three main research questions: 1) Were there clear variations in the implementation of teacher evaluation policies across Michigan school districts? 2) Which factors potentially affected such implementation decisions? and 3) How did the implementation of teacher evaluation policies affect student achievement? In this section, I review previous literature on factors that affect districts’ decision making and implementation of policies and the effects of teacher evaluation policies.

Districts’ Decision Making and Implementation of Policies

As Spillane (1996) pointed out, school districts received relatively little attention from policymakers for many years as most school reforms targeted either the state level or the local school level. Accordingly, school districts often appeared as background factors in policy implementation studies, rather than main agents for creating or implementing policies; they sometimes supported the enactment of policy from state to local schools, while they were frequently seen as barriers to this enactment (Honig, 2009). However, this trend has changed as school districts became one of the main participants in some

educational reforms (McLaughlin & Talbert, 2003). In particular, as tremendous amounts of data about teaching and learning became available along with multiple initiatives, including *NCLB*, “(f)ederal policies currently place unprecedented demands on school district central offices to use a range of sources of ‘evidence,’ ‘data,’ and ‘research’ to ground a host of decisions related to how central offices operate and how they work with schools” (Honig & Coburn, 2008, p. 580). That is, the role of school districts in making decisions about policies and instruction has become much greater. In turn, many research studies have examined the processes and effects of data-driven decision making at the district level (e.g., Carlson, Borman, & Robinson, 2011; Coburn, Toure, & Yamashita, 2009; Dembosky, Pane, Barney, & Christina, 2006; Honig & Coburn, 2008; Marsh et al., 2005; Park, & Datnow, 2009; Wohlstetter, Datnow, & Park, 2008).

In the teacher evaluation setting, Michigan school districts are both implementers and policymakers. A teacher evaluation policy framework was established at the state level, so they were implementers in this sense. However, individual school districts were left to make decisions about the details of the policies; for example, which observation tools and student growth measures they would use. That is, while they were implementers, they did not just follow prescribed roles, but they devised their own roles as they implemented the policies. Although their central role seems clear, it is not clear whether their decision making process in relation to teacher evaluation policy involved any “data.” Data for deciding how to implement the policies might not simply refer to student achievement data or data on effects of certain programs. Rather, data regarding the overall cost and benefits of each component of teacher evaluation policies, different observation tools, or student growth measures might be essential. However, it is questionable whether Michigan school districts had access to these types of data when they made decisions about the policies. Some have argued that many districts did not even have personnel skilled in different aspects of teacher evaluation policies (Keesler & Howe, 2016). Thus, I focused on reviewing studies that investigated factors affecting districts’ decision making and implementation of policies in general, as opposed to exclusively focusing on recent studies of school districts, including studies of data-driven decision making at the district level.

Resources are one of the most frequently referred to conditions for districts' decision making and policy implementation in general, as well as teacher evaluation settings. Without sufficient resources, such as time and fiscal resources, policies tend to stay at the surface level or even not be implemented (Coburn et al., 2009; Colby, Bradshaw, & Joyner, 2002; Donaldson, Woulfin, & Cobb, 2016; Kraft & Gilmour, 2016; McLaughlin, 1987; Spain, 2016; Steinberg & Donaldson, 2016). Implementing teacher evaluation policies in Michigan school districts in particular required a considerable amount of fiscal resources. Without additional support from the state, many districts needed to spend district funds to purchase teacher evaluation tools as well as student growth measures. This means that some districts with limited budgets needed to curtail other initiatives and reallocate resources to teacher evaluation. This motivates me to examine fiscal resources in each district as one of the potential factors that affect districts' decision making regarding teacher evaluation policies.

Another important factor for this process is different aspects of district leadership, including district administrators' expertise (Coburn et al., 2009; Spain, 2016) and buy-in to policies (Dutro, Fisk, Koch, Roop, & Wixson, 2002). That is, districts implement policies more actively when district administrators have the expertise to understand the core idea of policies, when they agree with the policies, and when they are able to devise a proper implementation plan accordingly. In this study, I focus on whether superintendents were members of a committee of a state-level organization, the Michigan Association of School Administrators (MASA), as a proxy for those aspects of leadership for two reasons. First, we can assume that those who served on the MASA committee were relatively experienced and respected leaders among district administrators. Second, they may have had stronger connections with the state-level government, given their representative status, which could help them understand the state system better. In addition, this connection may have contributed to a tight coupling between the state and their districts.

Student composition and district location can also influence districts' decision making regarding teacher evaluation policies. In a classic study of conflict management at the school district level, Boyd and Wheaton (1983) showed a clear difference between urban and suburban school districts. With limited

management resources, the decision making process in urban school districts featured “(a) much less data collection and analysis; (b) little use of expertise; (c) fewer committees and less citizen involvement; (d) fewer attempts to clarify or set goals and criteria; (e) less public discussion, more inclination toward secrecy and suspicion; (f) more inclination toward ad hoc and accidental policy development; and (g) more over politics and bargaining” (Boyd & Wheaton, 1983, p. 27). More recently, Neely (2015) argued that under *NCLB*, school districts that need to heavily rely on federal funding have faced a drastic increase in their administrative costs in order to maintain such revenue. Thus, it is important to consider districts’ demographics in teacher evaluation policy settings.

In addition to location and student composition, student achievement might play a significant role in districts’ decision making processes, due to the unique nature of teacher evaluation policies. Michigan school districts are evaluated annually based on student achievement on the MEAP (a state-wide standardized test), and their performance is publicly available in the form of Michigan School Scorecards. That is, raising student achievement scores is one of the most urgent issues among school districts and, as a result, teacher evaluation policies might have been implemented earlier or more rigorously in those districts that faced greater pressure to improve student achievement. In sum, based on previous literature, I examine how districts’ fiscal resources, location, and student composition, and student achievement scores seemed to affect their decision making regarding teacher evaluation policies.

The Effects of Teacher Evaluation Policies

As noted earlier, there have been few research studies regarding the effects of teacher evaluation policies as they are actually implemented at the district level. Rather, many research studies have been conducted at the individual teacher level or the school level with one or two components of teacher evaluation, often focusing on the reliability and validity of each component. In the first part of this section, I briefly review research on issues related to the reliability and validity of teacher evaluation instruments, and then turn to studies about the effects of teacher evaluation policies on students’ learning.

As accountability policies have shifted the target from the school level to the individual teacher level (Goldhaber, 2015; Lavigne, 2014), different tools for evaluating teachers have been rigorously

researched. In contrast with traditional evaluation systems, which have been criticized for failing to differentiate between degrees of teaching quality or to provide proper feedback for teachers, new teacher evaluation systems feature a “clearly defined set of teacher performance standards and/or a framework that defines good teaching and related indicators; rely on student learning gains as a significant factor in teacher performance ratings...” (Youngs & Haslam, 2012, pp. 1-2). Since these two components of teacher evaluation (i.e., observations using newly developed tools and students’ learning gains) comprise the majority of teacher evaluation systems in the U.S., many research studies have focused on them.

In particular, using data on students’ achievement gains, commonly known as value-added measures, has emerged as a popular way to evaluate teachers. The failure of conventional representations of teacher quality, such as advanced degrees or professional development, to predict students’ learning gains in a teacher’s classroom has partially contributed to this trend (Taylor & Tyler, 2012). However, while it seems straightforward to evaluate teachers based on students’ achievement scores using statistical techniques, estimating teachers’ value-added measures and using them as a significant component of teacher evaluation raises a number of issues. As Goldhaber (2015) noted, “there is not currently a consensus, or anything close to one, in the research community on the use of value-added measures for evaluation and decision making” (p. 87). Criticism of value-added measures ranges from technical issues of estimation, such as whether students are randomly assigned to teachers, comparability of units, and instability of the measures when only one or two years of data for teachers were available, to impacts on teachers’ instruction, such as encouraging teachers to “teach to the test” and a lack of useful feedback for teachers to enhance their instruction (Briggs & Domingue, 2011; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2013; Hallinger et al., 2014; Hill, Kapitula, & Umland, 2011; Rothstein, 2010). Based on these concerns, researchers have studied teacher value-added measures in conjunction with other observation instruments.

Research studies using both teachers’ observation data and value-added measures provide meaningful insights not only into teacher evaluation policies but also into measuring teacher quality itself. Researchers have reported low to moderate levels of correlation between these two measures of teachers’

quality in general. The MET study (2012) reported correlations between teachers' observation scores and value-added measures ranging from 0.12 to 0.34. Similarly, Milanowski (2004) reported a 0.27 correlation coefficient between two measures of teaching quality, and Rockoff and Speroni (2010) reported correlations ranging from 0.21 to 0.26. These weak correlations between the two measures cast doubt on the notion of a single dimension of effective teaching (Harris et al., 2014; Rothstein & Mathis, 2013). That is, effective teachers based on two measures can differ: in Harris and colleagues (2014), principals rated some high value-added teachers as ineffective because those teachers tended to be isolated from other teachers. However, it is interesting that, even if one only cares about students' achievement as a result, principals' observation can provide useful information for teachers' effectiveness (Kane, Taylor, Tyler, & Wooten, 2011). Based on these findings, it can be argued that different measures of teaching quality can capture different aspects of effective teaching, so it is necessary to use multiple measures to capture teachers' instruction, which is consistent with current teacher evaluation policy systems in many jurisdictions in the U.S.

Now I turn to studies that directly examined the effect of teacher evaluation policies on student learning. As noted above, in contrast with a large body of research studies on the reliability and validity of each component of teacher evaluation, the effects of policy enactment on students' learning have not yet been extensively studied. Moreover, the existing research has shown mixed results.

Some researchers are skeptical about the effects of teacher evaluation policies for various reasons. In an analytic essay, Lavigne (2014) pointed out some limitations of the three business performance evaluation models that Race to the Top was following: the rating scale method, the ranking method, and the forced distribution method. Based on examples in the business sector, Lavigne (2014) concluded that current teacher evaluation policies would not improve students' learning in the long run. Murphy, Hallinger, and Heck (2013) made a similar argument supported by the fact that a well-established body of studies about school improvement rarely included teacher evaluation as an important method. Kraft and Gilmour (2016) pointed out the difficulties of implementing teacher evaluation policies. The authors argued that the reliability of the newly developed teacher evaluation instruments cannot guarantee that

new policies indeed lead to changes in teachers' instructional practice, since evaluators tend to make conscious choices about teacher evaluation ratings considering issues other than the pure quality of teaching, such as their degree of comfort with the teacher being evaluated.

In contrast, two empirical studies suggested a positive effect of teacher evaluation policies on student learning. Taylor and Tyler (2012) showed that students whose teacher was evaluated by their school district had mathematics scores that were 0.10 standard deviation higher than similar students taught by the same teacher before the teacher was evaluated. Since teachers in their study could not self-select the year of their evaluation, as the district required teachers to be evaluated every fifth year after they started teaching, the authors were able to compare the same teachers before and after being evaluated. However, it should be noted that the evaluation system featured in the study, which was in place in Cincinnati between 2003-04 and 2009-10, is quite different than the current systems in place in many districts and states. First, the system only included classroom observations by principals and peer teachers. Second, teachers were evaluated only once in every five years in this system. Third, the results from the evaluation were not used for high-stakes decisions regarding teachers' job status.

The second study is based on a pilot study of a teacher evaluation system for the Chicago Public Schools from 2008 to 2010 (Steinberg & Sartain, 2015). The authors conducted an experiment to examine the effects of a teacher evaluation system featuring classroom observations. As in the previous study, the system studied did not include student achievement. In order to claim a causal relationship between enactment of the policies and student achievement, the authors randomly assigned treatment and control groups at the school level and enacted the system only for treatment schools. Although there was no significant difference between the treatment and control groups with regard to math achievement, the authors found significant effects of the policy on reading achievement. There are three limitations to be noted in this study: 1) the unit of analysis was at the school level, while it is more likely that districts will enact new teacher evaluation policies; 2) the study allocated extensive resources for training principals, but this might not be realistic for many districts with limited resources; and 3) the study was conducted as a pilot, so no high-stakes decisions were attached.

Taken together, despite some empirical evidence of positive effects of the policies and some concerns about them, it is unclear how student achievement is affected by teacher evaluation policies at the district level, which often include multiple measures of teaching quality and high-stakes decisions. The current study addresses this gap in the literature by analyzing the effects of teacher evaluation policies as they have actually been implemented.

Michigan Teacher Evaluation Policies

While teacher evaluation was almost entirely under local control in Michigan prior to 2010, this significantly changed as Public Act No. 102 of 2011 created a statewide system of teacher evaluation (Michigan Department of Education, 2014; Pogodzinski et al., 2015). The current legislation requires all teachers and administrators to be evaluated at least annually with a “rigorous, transparent, and fair evaluation system” (Mich. Comp. Laws § 380.1249). It emphasizes using multiple rating categories (i.e., highly effective, effective, minimally effective, and ineffective) and including student growth as a significant factor. In 2014-15, teachers whose grades and subjects were included in the state standardized assessment were required to use these scores (i.e., a student growth measure) for at least part of their evaluation. Although the legislation required at least 25% of a teacher’s evaluation to be based on student growth, the Michigan Senate recently approved legislation that lowers the proportion of students’ growth in teachers’ summative evaluations (Oosting, 2015). Teacher evaluation is excluded from collective bargaining in Michigan (Public Employment Relations Act, Mich. Comp. Laws § 423.215 2014). [See Appendix A for the details of the state law]

Despite this state-level law, there has been a high level of discretion among Michigan school districts; Steinberg and Donaldson (2016) categorized Michigan as “state-designed system with local discretion” (p. 6), and Gagnon, Hall, and Marion (2016) also categorized the state as a high local control state, based on its evaluation procedures for teachers in non-tested subjects and grades. MDE requires teachers of non-tested grades and subjects to include student growth measures as a major component of teacher evaluation. In particular, the growth measures need to be “multiple research-based growth measures or alternative assessments that are rigorous and comparable across schools within the

school district, intermediate school district, or public school academy (380.1249(2)(a)(ii))” (Michigan Department of Education, n.d., p. 12). Again, however, the details of how those teachers are evaluated depend on each district.

Such a high level of local control is manifested in three ways. First, even though it is noted in the law that the results of the evaluation are used in determining teacher retention, promotion, and termination, how each district should use evaluation information is not specified. Second, there is no regulation regarding classroom observation tools or student growth measure. Districts can use or modify the evaluation tools that the state recommends, but it is also possible for them to use other tools that districts create themselves. Third, the state system does not set clear guidelines for categorizing teachers as highly effective, effective, minimally effective, or ineffective, resulting in rather ambiguous and subjective rating criteria (Pogodzinski et al., 2015). This high level of autonomy can be explained in part by the fact that the state was mandated to enact the policies, but there was no financial support from the federal government. Michigan received a waiver resulting from *NCLB* that required the implementation of teacher evaluation policies, while it failed to win federal funds (i.e., Race to the Top) to support the policies (Keesler & Howe, 2016). With such limited resources, MDE failed to provide systematic support for school districts, and at the same time, it did not have authority to enforce the implementation of teacher evaluation policies (Keesler & Howe, 2016).

In fact, there are significant variations among Michigan districts. According to a report by MDE (2014), while most districts included instructional practice as a teacher evaluation component, only half of the districts used student growth measures to evaluate their teachers during the 2013-14 school year. Moreover, 44% of districts used locally developed tools to evaluate teachers’ instructional practice. The weight given to student growth also varies; even though the state requires that it account for more than 25% of a teacher’s evaluation rating as of 2013-14 school year, about 10% of districts reported that it made up less than 20% of teachers’ ratings. In terms of the use of teacher evaluation ratings, most districts reported that they used the results to target professional development, while only 60% of districts used the information for termination. However, it is interesting to note that only 3% of Michigan teachers were

rated as minimally effective or ineffective in 2013-14. That is, it is not clear how districts used teacher evaluation results for such decisions given the lack of variation in those results.

Method

Data

The goals of this essay are to investigate 1) variation in the implementation of teacher evaluation policies, specifically in the timing of the implementation and teacher dismissal across Michigan school districts; 2) factors that might affect such decisions at the district level; and 3) the effects of teacher evaluation policies on student achievement.

For the main variable of interest, variation in policy implementation, I administered a survey to district administrators. I contacted 350 district administrators in 179 school districts in Michigan in 2015-16. I had two criteria for selecting school districts: 1) the school district served more than 2,500 students; and 2) the school district was known to use Danielson Framework for Teaching, according to Pogodzinski and colleagues' data (2015). The second criterion was used because the Danielson Framework was one of the most common teacher evaluation tools in Michigan, and it was helpful to include districts that shared the same tool when it came to estimating the effects of the policies. As a result, 13 relatively small school districts were added to the sample. However, during data collection, I learned that many districts had changed their observation tool multiple times, and there was no dominant evaluation tool used by districts. Thus, it would be less meaningful to analyze the effect of the policies by different evaluation tools, so I do not consider the tools in the analysis. In addition, since charter schools were subject to different regulations than traditional public school districts, making it hard to compare them to school districts, I only focused on traditional public school districts. Out of the 179 sampled school districts, administrators in 101 school districts completed the survey. In some cases, two administrators working at the same district completed the survey and their responses were not identical; in this case, I used the response from the person who had worked at the district for a longer period of time.

It is important to note a self-selection issue in this study — districts that did not complete the survey might have some reasons for not participating, meaning that missing values are not randomly

distributed. Implementing teacher evaluation involves legal issues, given that the state required districts to enact the policies at a certain time point. From districts' perspectives, they may be reluctant to complete a survey that asks about the timing of their enactment of the policies if they believe that they did not fully comply with the law. To address this, in my initial contact with districts, I clearly indicated that I was affiliated with Michigan State University and my focus was examining the phenomenon of teacher evaluation, rather than evaluating districts' compliance with the state law. In the consent form, I also stated that I would not share identifiable data with anyone outside of this project. Nevertheless, variation regarding implementation of the policies derived from the survey can be considered as a lower bound of the variation in policy implementation across all of Michigan. In terms of comparisons between respondents and non-respondents, districts that completed the survey tended to serve a higher proportion of proficient students in mathematics and reading and higher percentages of White students. However, there were no statistically significant differences between the two groups in their mean scale scores of mathematics and reading, the standard deviation of their scores, their locations (i.e., suburban, urban, or rural), their total levels of enrollment, or the proportion of free and reduced lunch eligible students that they served.

Along with the survey, I interviewed 11 school districts administrators and two representatives of the Michigan Education Association. Except for two school district administrators and the MEA representatives, all interview participants completed the survey prior to the interview. In general, interviews lasted about an hour, and the interview questions focused on details about districts' decision making processes with regard to the implementation of teacher evaluation policies and factors affecting such processes [See Appendix B for the interview protocol]. Although a complete analysis of these qualitative data is beyond the scope of this essay, the interview data informed the identification of variables in the survey data that could explain variations in district implementation of teacher evaluation policies.

In order to answer the second and third research questions, I drew on multiple publicly available data sets. In terms of student achievement, I drew on MEAP (Michigan Educational Assessment Program)

test results in mathematics and reading from the 2007-08 to 2013-14 school years at the district level. Michigan enacted a new test system, the M-STEP (Michigan Student Test of Educational Progress) in the 2014-15 school year. The M-STEP is “a very different test than tests administered in past years, therefore, results should not be compared to those from prior years” (Michigan Department of Education, n.d.) and, thus, I did not include test results after the 2013-14 school year. In terms of factors that might affect districts’ decision making regarding teacher evaluation policies, I collected data about district demographics (i.e., location, total enrollment, proportion of free and reduced lunch eligible students, and proportion of White students) and the total revenue of each school district from the MDE website. MASA committee membership data were derived from the MASA website. Table 10 presents the descriptive statistics for the school districts in the study.

Table 10. Descriptive Statistics for School Districts

	M	SD
District total revenue in 2007-08 (log)	17.803	0.683
MASA committee membership in 2013	0.069	-
Student achievement in Mathematics in 2007-08		
Proportion of proficient students (%)	38.103	13.361
Mean scaled score	575.346	9.253
Standard deviation	25.494	2.506
Student achievement in Reading in 2007-08		
Proportion of proficient students (%)	63.094	10.997
Mean scaled score	578.983	8.116
Standard deviation	28.45	1.018
Student composition in 2007-08		
Proportion of White students (%)	81.661	17.193
Proportion of free and reduced lunch eligible students (%)	28.569	17.114
Total enrollment	5388.356	3607.966
District location		
Suburban	0.584	-
Urban	0.33	-
Rural	0.089	-

Note. Models about factors that might affect the implementation of teacher evaluation policies (RQ2) included all time-invariant variables listed above, while models about the effects of teacher evaluation policies (RQ3) included student achievement and composition variables as time-variant variables from the 2007-08 to 2013-14 school year data. Suburban, urban, rural, and MASA committee membership are dichotomous variables. N=101.

Measures

The implementation of teacher evaluation policies. The survey items about the implementation of teacher evaluation policies include the following: When did your district . . . 1) evaluate all teachers, including probationary and tenured teachers for the first time? 2) evaluate all teachers, including probationary and tenured teachers on an annual basis for the first time; 3) use teacher evaluations to make decisions about teacher financial incentives for the first time; 4) use teacher evaluations to make decisions about teacher promotion for the first time; 5) use teacher evaluations to make decisions about teaching assignment for the first time; 6) use teacher evaluations to make decisions about transfer for the first time; and 7) use teacher evaluations to make decisions about dismissal for the first time. Respondents answered these items by indicating school years (i.e., 2016-17 year) or Not Applicable if they had never implemented a given aspect of the policies. Based on the law, I defined the timing of the implementation for each district based on two steps. First, I took the earliest year from the responses for items 3 through 7. That is, I treated the use of teacher evaluation results in any of these ways as part of implementing the policies. Second, I took the latest year from the responses for items 1 and 2, and the year calculated from the first step. In sum, I took the school year when each district first evaluated all teachers annually and used teacher evaluation results for any of the listed purposes as the year of implementation. This implementation year variable was included in the models investigating the third research question about the effects of teacher evaluation policies on student achievement.

For the second research question, I created four dichotomous variables based on the year of implementation variable. Since the state enacted the policies in the 2011-12 school year, districts that implemented the policies before the 2011-12 school year were coded as early adopters (i.e., 1=early adopters and 0=else); districts that implemented the policies after the 2012-13 school year¹ were coded as late adopters (i.e., 1=late adopters and 0=else); and non-compliers were the districts that never implemented the policies as required (i.e., 1=non-compliers and 0=compliers). For example, districts that implemented the policies in the 2011-12 school year had a value of 0 for all three dummy variables, since they implemented the policies on time. In terms of teacher dismissal, if a district administrator indicated

any school year for question 7, I coded it as a dismissal adopter (i.e., 1=dismissal adopters and 0=else). These variables were included separately in the models.

Factors that might affect the implementation of the policies. Districts' total revenue is from the 2007-08 year data, and it is a sum of all of the funds for districts from different sources, including Title 1. I took the natural log of these values in order to create a normal distribution. Being a member of a state-level committee is a dichotomous variable equal to 1 if a district's administrator was a member of the systemic school reform committee of MASA in 2013. The data before 2013 were not available. In addition, I included the proportion of proficient students and the standard deviations of students' scores in mathematics and reading, the proportion of White students and free- and reduced-lunch eligible students, total enrollment, and dummy variables for suburban and rural districts. This district demographic information is derived from 2007-08 school year data.

Student achievement. The main form of the student achievement variables is the proportion of 3rd- to 8th-grade students at the district level who were proficient on state tests in mathematics and reading. This might be more reliable than the mean scaled scores given that mean scaled scores can be heavily influenced by the test itself across different years. However, as a robustness check, I used mean scaled scores as well. For the main models, I used aggregated data from all students, while I also analyzed the effects of the policies on different groups of students separately, such as female/male, economically disadvantaged/non-disadvantaged, White/Black/Hispanic, and elementary/middle school students. Since student achievement in science and social studies is only available for 5th- and 8th-graders, I only focused on mathematics and reading scores.

Analytic Approach

The first research question focuses on variation in the timing of enactment of the policies and whether districts used teacher evaluation results for decisions regarding teacher dismissal. The second question addresses factors that might affect such decisions. I used logistic regression for this analysis.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{Proficient_student}_i) + \beta_2(\text{SD}_i) + \beta_3(\text{Pr_White}_i) + \beta_4(\text{Totalenrollment}_i) + \beta_5(\text{Suburban}_i) \\ + \beta_6(\text{Rural}_i) + \beta_7(\text{Pr_FRL}_i) + \beta_8(\text{Log_Rev}_i) + \beta_9(\text{MASA}_i) + e_i \quad (1)$$

Where p is the possibility that a district is an early adopter, late adopter, non-complier, or dismissal adopter. These outcome variables were included in separate models. $\text{Proficient_student}_i$ is the proportion of students who were proficient in mathematics in 2007-08 school year², and SD_i is the standard deviation of mean scaled scores on the 2007-08 test. Pr_White_i , Totalenrollment_i , and Pr_FRL_i are the proportion of White students, total enrollment of the district, and the proportion of free- and reduced- lunch eligible students in the 2007-08 school year, respectively. Log_Rev_i is a natural log of total revenue of school districts in the 2007-08 school year; MASA_i is a dummy variable equal to 1 if a district's superintendent participated in the MASA systemic school reform committee in 2013. All variables in this model are at the district level. Since districts in the same intermediate school districts (ISD) can influence each other, I used cluster robust standard errors at the ISD level. To check whether student achievement had a non-linear association with the outcomes, I included a dummy variable equal to 1 if a district fell into the lowest quartile in terms of the proportion of students with proficient scores, instead of linear terms of the variables. Using this dummy variable did not lead to much difference in the results. Although a dummy variable regarding whether a district had at least one priority school (i.e., student achievement was among the lowest 5% of all Michigan schools) was considered, it was not feasible to run these models because only two districts in the data had a priority school in 2010.

It is important to note that some districts were excluded from some models based on when they implemented the policies. In terms of the analysis of the early adopters, districts that implemented the policies at some point prior to the 2008-09 school year were excluded. Student achievement in the 2007-08 school year, which is the earliest year of data among all available data, can be a result of the policies, and it is also hard to assume that policies implemented prior to 2008 were similar to the state system enacted in 2011-12. For the analysis of the late adopters, non-compliers were excluded, since they had never implemented the policies as required. However, these non-compliers were included in the analysis

of dismissal adopters as some of them implemented this particular aspect of the policies while others did not. Again, it is not that non-compliers never implemented any parts of the policies; they might have implemented some aspects of the policies, but they did not enact all of them. For some models, I included reading scores, instead of mathematics scores, and different forms of achievement scores (i.e., mean scaled scores) along with other district demographic information, but the results mostly stayed the same. In order to check whether there was a significant multicollinearity issue, I included some variables that theoretically related to each other separately before I proceeded to the main model.

In order to examine the third research question about the effects of teacher evaluation policies on student achievement, I applied an ITS model. Before applying this model to establish a causal relationship between implementation of teacher evaluation policies and student achievement, there were three main assumptions that needed to be considered. First, the trend of the outcome variable, student achievement scores, needs to be linear, so that the data from pre-treatment years can be a plausible counterfactual for the post-treatment years. The proportion of proficient students at the district level from 2007-08 to 2013-14 was generally linear, especially conditional on proportion of students eligible for free- and reduced-lunch at each district. Second, the timing of the implementation of the policies is randomly assigned. This is a very important assumption for correctly estimating the pure effects of the policies; if there are unobserved factors that affect both the timing of the policies and student achievement, the estimate would be biased. The results from the second research question might be useful for this aspect, since they can explain some of the factors that affected the timing of the policies. Accordingly, I included the variables used for model (1), except for the student achievement, in the models investigating the third research question. Third, there should be no concurrent event at the same time that the teacher evaluation policies were enacted. I asked this question during the interviews with district administrators. Some districts had some district-level initiatives (e.g., new reading curriculum, etc.), but there was no systemic change for multiple school districts in general. Model (2) describes a modified version of ITS model.

$$Y_{it} = \beta_0 + \beta_1(\text{Year}_{it}) + \beta_2(\text{Policy}_{it}) + \beta_3(\text{Year_Since_Policy}_{it}) + \mathbf{X}_{it}\boldsymbol{\gamma} + u_i + e_{it} \quad (2)$$

Where, Y_{it} is the proportion of students with proficient MEAP test scores for district i in year t , which are mathematics and reading in separate models. In addition to drawing on the achievement data for all students, I also used the achievement data from different students for some models, such as female/male, economically disadvantaged/non-disadvantaged, White/Black/Hispanic, and elementary/middle school students. Year_{it} is year linear term, centered at the 2007-08 school year. Policy_{it} is defined as 0 if the policies are not enacted at time t in the district i , as 1 if the policies have been enacted. $\text{Year_Since_Policy}_{it}$ represents the post-policy trend; it is equal to 0 until a district enacts the policies and equal to 1 after one year of implementation, equal to 2 after two years of implementation, and so on. $\mathbf{X}_{it}$ is a vector of time-varying covariates within districts such as the proportion of White students and free-and reduced-lunch eligible students, and total enrollment. u_i represents district fixed effects and e_{it} is a random error with mean of zero. The main interest is β_2 and β_3 ; β_2 captures the immediate effects of the policies after controlling for the trend in student achievement and district characteristics, while β_3 captures the effects of the policies on the slope of the student achievement after the policies. For this ITS model, I excluded non-compliers.

There are two main reasons why an ITS model is more appropriate than a comparative interrupted time series (CITS) model for this study. First, the pre-treatment trend of the outcome of control groups, non-compliers in this case, was not linear. Thus, including non-compliers as a comparison group does not increase the robustness of the analysis. Second, since non-compliers never implemented the policies as required, there were no post-treatment years for this group. For example, in Dee and Jacob (2011), there was a certain time point (i.e., 2002-03 school year) that all states implemented NCLB, the treatment. However, in the current study, there was no certain time point that all districts in Michigan were subject to the same regulation. Moreover, many compliers enacted the policies at different times. To be sure, although it is possible to establish a time of enactment such as the 2011-12 school year when MDE required all districts to enact the policies, this introduces some uncertainty to the analysis. In addition,

there were only eight non-compliers in the data. Accordingly, I applied the ITS model noted above as the main model, but conducted a CITS model as a secondary analysis. Model (3) describes a modified version of the CITS model.

$$Y_{it} = \beta_0 + \beta_1(\text{Year}_t) + \beta_2(\text{Policy}_{it}) + \beta_3(T_i * \text{Year}_t) + \beta_4(T_i * \text{Policy}_{it}) + \beta_5(\text{Year}_t * \text{Policy}_{it}) + \beta_6(\text{Year}_t * T_i * \text{Policy}_{it}) + \beta_7 X_{it} + u_i + e_{it} \quad (3)$$

Where, T_i is equal to 1 if a district has ever implemented the policies and 0 if a district has never implemented the policies as required, and other parts stay the same.

In order to understand the results from these analyses, it is important to note how the comparison works for these analyses. For the ITS model, the counterfactual is each district's pre-treatment trend. As noted above, most districts implemented the policies gradually. In those cases, the model estimates the effects of full implementation in relation to partial implementation. This applies to the CITS model as well. There were eight control group districts, and they implemented aspects of the policy reform to varying degree.

To be specific, three districts were coded as the control group because they had never used teacher evaluation results for any decisions related to teachers; two districts had never evaluated their teachers annually; two districts had never evaluated all of their teachers; and one district had never evaluated all teachers annually. That is, this analysis does not compare districts that fully implemented teacher evaluation policies versus those that never implemented them; instead, it compares districts that fully implemented the policies versus those that partially or slowly implemented them.

Lastly, in order to determine if some extreme outliers, in terms of student achievement, led the results, I ran the models noted above with and without 10 extreme districts (i.e., five districts that served the most proficient students in mathematics and five districts that served the least proficient students in mathematics). The results stayed almost identical and, thus, I focused on the results using all available data for the following sections.

Results

Variations in the Implementation of Teacher Evaluation Policies

I start the analysis by investigating two types of variation in districts' implementation of teacher evaluation policies: 1) timing of implementation and 2) whether they used teacher evaluation ratings for decisions about teacher dismissal. As Table 11 shows, there is notable variation in both aspects. Although most districts (49 school districts) implemented the policies in the 2011-12 or 2012-13 school years as the law required, there were 36 districts that enacted the policies later or still had not implemented the policies as of 2015-16. In terms of using teacher evaluation ratings for decisions about teacher dismissal, it is surprising that the mode year was prior to 2008 (34 school districts), while a considerable number of districts (17 school districts) had never used teacher evaluation ratings to make decisions about teacher dismissal. This point stood out during the interviews as well. A Human Resource (HR) director working at District A stated that they had never quantified teachers' ratings or used teacher evaluation ratings for teacher dismissal because it undermined their culture of professional learning communities. In contrast, a HR director working at District B indicated that they ranked teachers from top to bottom, and dismissed teachers from the bottom. In this situation, the law did not lead to a distinct change in District A, in terms of how teacher evaluation ratings were used, while it might have led to changes in District B. Overall, with regard to the implementation of teacher evaluation policies, the system between school districts and the state government in Michigan still seemed to be loosely coupled both in terms of distinctiveness and responsiveness. Districts A and B were distinctive in their ways of using teacher evaluation ratings, but both of them were responsive to the state law.

Table 11. The Timing of the Enactment of Teacher Evaluation Policies

	Year of implementation	Year of using teacher evaluation ratings for teacher dismissal
Prior 2008	6	34
2008-09	0	0
2009-10	1	0
2010-11	9	6
2011-12	30	9
2012-13	19	9
2013-14	18	10
2014-15	8	11
2015-16	2	5
N/A(never implemented)	8	17

Note. Year of implementation is the first year that each district enacted the policies as required for the first time (evaluating all teachers on annual basis and using teacher evaluation ratings for making decisions about teachers). Year of using teacher evaluation ratings for teacher dismissal is the year that each district used teacher evaluation ratings to make a decision about teacher dismissal for the first time. N=101.

Factors that Might Affect the Implementation of Teacher Evaluation Policies

In order to examine the second research question, I conducted logistic regression models on four separate dichotomous variables: districts being early adopters, late adopters, non-compliers, and dismissal adopters. Based on the survey data, I coded 16 school districts as early adopters; 28 school districts as late adopters; 8 school districts as non-compliers; and 84 school districts as dismissal adopters. The potential factors included in the models were student achievement, student composition in terms of race and socio-economic status, superintendent's membership on a state-level committee, and district's total revenue. As noted above, theoretically, some variables can be highly correlated with one another, such as student achievement, the proportion of White students, the proportion of free- and reduced-lunch eligible students, and district location. Thus, I included those variables one-by-one for models estimating the association between districts being early adopters and other factors before I proceeded to the main analysis [See Appendix C for the details]. Although the standard errors for some variables increased as multiple variables about district background were entered into the same model, the magnitude of the changes in the standard errors was not significant, and the statistical inferences largely stayed the same.

Table 12 reports the results from the logistic regression on four outcome variables (i.e., early adopters, later adopters, non-compliers, and dismissal adopters), using proportion of proficient students/mean scaled scores in mathematics/reading MEAP tests in 2007-08 along with other factors. First, early adopters were active in the area of teacher evaluation and took anticipatory actions even before the law was enacted. Since none of the MASA committee districts were early adopters of the policies, this variable was dropped for this part of analysis. Proportion of White students and total revenue of districts had significant positive associations with the odds of being an early adopter of the policies; based on Model 1, having 10% more White students at the district level is associated with 11 times higher odds of being an early adopter of the policies; 10% more district revenue is associated with 1.8 times higher odds of being an early adopter of the policies, after controlling for other district characteristics. Neither using mean scaled scores nor reading achievement scores led to a significant change in the results.

The result involving district revenue can be explained in part by interview data. Most interview participants noted that they were obligated to purchase a teacher evaluation tool, such as the Danielson Framework for Teaching (Danielson, 1996) or the Marzano Causal Teacher Evaluation Model (Marzano, Toth & Schooling, 2012), as well as a tool for measuring student growth, such as Northwest Evaluation Association (NWEA) student assessments, without financial support from the state. It is very hard for districts to craft a new evaluation framework as well as provide training for evaluators and professional development for teachers, and create a system for reporting and storing the data by themselves. Thus, purchasing a package of teacher evaluation tools that included all of the above was a more realistic option for districts. In terms of student growth measures, although the state indicated that state-level standardized tests, such as MEAP and M-STEP tests could be used, districts typically did not have access to these data at the time when teacher evaluations were supposed to be completed. In addition, most district administrators acknowledged that such state-level tests were not reliable because the results were not stable across years. Therefore, the fiscal resources available at the district level might be critical in terms of the implementation of teacher evaluation policies in each district. This might be the case for other

Table 12. Factors that Might Affect the Implementation of Teacher Evaluation Policies

<i>Early Adopters</i>	Mathematics achievement		Reading achievement	
	Model (1)	Model (2)	Model (3)	Model (4)
Proportion of proficient students	0.958 (0.0779)	-	1.019 (0.0825)	-
Standard deviation	1.228 (0.46)	1.376 (0.511)	1.574 (0.677)	1.551 (0.666)
Mean scaled score	-	0.892 (0.109)	-	1.013 (0.106)
Total enrollment	0.999 (0.000437)	0.999 (0.000426)	0.999 (0.000334)	0.999 (0.000335)
Proportion of White students	1.118** (0.0460)	1.127** (0.041)	1.104** (0.0408)	1.105** (0.0408)
Proportion of free and reduced lunch eligible students	1.028 (0.0558)	1.015 (0.0549)	1.047 (0.0768)	1.042 (0.0759)
District total revenue (log)	18.86* (25.51)	20.59* (27.29)	11.60* (14.14)	11.95* (14.84)
Suburban	1.355 (1.050)	1.316 (1.025)	1.515 (1.422)	1.456 (1.32)
Rural	2.121 (1.943)	2.037 (1.828)	2.672 (2.73)	2.631 (2.698)
N	95	95	95	95
<i>Late Adopters</i>	Mathematics achievement		Reading achievement	
	Model (1)	Model (2)	Model (3)	Model (4)
Proportion of proficient students	1.002 (0.0486)	-	0.983 (0.0323)	-
Standard deviation	0.994 (0.197)	0.978 (0.0616)	1.12 (0.266)	1.133 (0.273)
Mean scaled score	-	1.058 (0.191)	-	0.979 (0.0398)
Total enrollment	1 (0.000143)	1 (0.000142)	1 (0.000145)	1 (0.000145)
Proportion of White students	0.978 (0.0197)	0.98 (0.02)	0.981 (0.0194)	0.98 (0.0196)
Proportion of free and reduced lunch eligible students	0.985 (0.0255)	0.979 (0.0259)	0.975 (0.0281)	0.976 (0.0265)
District total revenue (log)	1.452 (1.304)	1.395 (1.265)	1.332 (1.208)	1.331 (1.212)
MASA committee	1.616 (1.674)	1.744 (1.798)	1.791 (1.906)	1.782 (1.899)
Suburban	3581319*** (2729993.6)	2214173.2*** (1724975.7)	2736043.9*** (2143221.2)	2774830.7*** (2166742.1)
Urban	2978087.6*** (2380245.7)	1801927*** (1405962.1)	2172886.3*** (1710310.1)	1714261*** (1714261)
N	93	93	93	93

Table 12 (cont'd)

<i>Non-compliers</i>	Mathematics achievement		Reading achievement	
	Model (1)	Model (2)	Model (3)	Model (4)
Proportion of proficient students	0.792** (0.0722)	-	0.809** (0.06)	-
Standard deviation	2.152* (0.755)	1.743 (0.56)	0.679 (0.282)	0.777 (0.318)
Mean scaled score	-	0.763* (0.095)	-	0.823* (0.074)
Total enrollment	1 (0.000243)	1 (0.000261)	1 (0.00024)	1 (0.00024)
Proportion of White students	1.098* (0.0448)	1.094 (0.052)	1.104 (0.0941)	1.091 (0.0856)
Proportion of free and reduced lunch eligible students	1.006 (0.0552)	1.007 (0.06)	0.96 (0.0601)	0.992 (0.0618)
District total revenue (log)	2.814 (3.374)	2.062 (2.707)	0.827 (1.064)	1.16 (1.527)
MASA committee	10.16 (21.11)	6.646 (14.29)	5.991 (13.25)	4.303 (9.282)
Suburban	1.681 (2.107)	1.27 (1.554)	0.502 (0.571)	0.618 (0.701)
Rural	26.68* (35.96)	16.71* (20.008)	7.619 (8.613)	8.925 (10.71)
N	101	101	101	101
<i>Dismissal Adopters</i>	Mathematics achievement		Reading achievement	
	Model (1)	Model (2)	Model (3)	Model (4)
Proportion of proficient students	1.053 (0.0604)	-	0.914 (0.0516)	-
Standard deviation	0.599* (0.163)	0.653 (0.169)	0.736 (0.28)	0.773 (0.29)
Mean scaled score	-	1.041 (0.0843)	-	0.88 (0.654)
Total enrollment	1 (0.000153)	1 (0.000142)	1 (0.00014)	1 (0.00015)
Proportion of White students	0.96 (0.0274)	0.963 (0.0262)	0.987 (0.0263)	0.986 (0.0257)
Proportion of free and reduced lunch eligible students	0.986 (0.0216)	0.98 (0.0241)	0.956 (0.0306)	0.953 (0.0293)
District total revenue (log)	3.381 (3.115)	3.486 (3.158)	2.652 (2.219)	2.638 (2.238)
MASA committee	0.0628* (0.0735)	0.0738* (0.0837)	0.112 (0.129)	0.115 (0.135)
Suburban	1.032 (0.764)	1.132 (0.795)	1.88 (1.289)	1.985 (1.35)
Rural	0.341 (0.308)	0.356 (0.321)	0.371 (0.349)	0.363 (0.346)
N	101	101	101	101

Table 12 (cont'd)

Note. Cluster robust errors at the ISD level were used. The student achievement variables used in these models are based on MEAP test scores in mathematics and reading in the 2007-08 school year. The coefficients are expressed in odds-ratio for the sake of easy interpretations. MASA committee, suburban, and urban are dichotomous variables. For the early adopter analysis, MASA committee variable was omitted because none of the MASA committee districts were early adopters; for the late adopter analysis, a dummy variable for urban was included instead of the rural variable, because no rural districts were late adopters in the data. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

reforms, but the difference in this case is that the state government did not provide any resources to implement the policies.

Against my expectation that districts that faced greater pressure to increase students' test scores might be early adopters of the policies, student achievement had no significant relationship with districts being early adopters of the policies. This might indicate that teacher evaluation policies were not identified as a promising tool for improving student achievement by district administrators. This point was also raised during the interviews; while many participants agreed that teacher evaluation policies can improve student learning, they were skeptical about the implementation process. The superintendent in District C noted, "I appreciate and respect the intent of the legislators when they set forth the provisions in law but the reality is that it just doesn't always match up as nicely as it could or should in their eyes." If district administrators perceive teacher evaluation policies as another requirement with which they need to comply, rather than a tool for improving teaching and learning, districts serving many low-achieving students would likely not be interested in teacher evaluation policies.

While late adopters of the policies implemented the policies after the deadline, they did eventually enact them. That is, they might have been very cautious about the policies and implemented them gradually as they witnessed changes in the state system and other districts' implementation. Most district characteristics did not have significant associations with districts being late adopters. Since there were no late adopters located in rural districts, it was not possible to estimate the effects of district location based on the data.

Non-compliers were districts that had never implemented the policies as required; they usually did not implement one or two aspects of the policies (e.g., evaluating all teachers annually or using

teacher evaluation ratings for making decisions for teacher related issues). A consistent pattern in the analysis was that having higher student achievement was associated with a lower odds of being a non-complier. It did not matter whether reading or mathematics achievement scores were used or whether proportion of proficient students or mean scaled scores were used. In Model (1), a 1% increase in the proportion of students who were proficient on the mathematics test is associated with about 20% less odds of being a non-complier, controlling for other district characteristics. In other words, districts serving many low-achieving students were less likely to implement the policies as required. This result is consistent with the one about early adopters; district administrators might not perceive teacher evaluation policies as a useful tool for enhancing student learning. Rather, it could be perceived as a burden for these districts due to the resources that they need to spend on the policies, such as time, fiscal resources, and human resources. In fact, during the interview, the superintendent in District D argued that it was the changes they made in curriculum that affected student learning, not their teacher evaluation policies. Rural school districts were more likely to be non-compliers when student achievement in mathematics was included, but this association was no longer significant when reading achievement was included.

Lastly, dismissal adopters were districts that implemented a specific part of the policies, using teacher evaluation ratings to make decisions about teacher dismissal, regardless of other components of the policies or the timing of implementation. As noted earlier, this part of teacher evaluation policies is particularly controversial; there might be a number of obstacles facing districts with regard to implementing this aspect of the policies. That is, dismissal adopters can be considered as another type of active implementers of the policies. There was no clear pattern in the results, but it is worthwhile to note that superintendents' MASA committee membership had a negative association with districts being dismissal adopters. In other words, the MASA committee members seemed to refuse to use teacher evaluation ratings for dismissal decisions. This is again opposite to my expectation that membership on this state-level committee might increase the responsiveness of those districts to state-level policies, so they are more likely to implement them. In fact, those districts were less responsive to this specific aspect of the policies. It is also important to note that this variable has never been significant in previous

analyses. However, since this association was no longer significant when reading achievement scores were inserted, and there were only seven MASA committee districts in the data, it is necessary to be cautious about this result.

Effects of Teacher Evaluation Policies on Student Achievement

The third research question is about the effects of the implementation of teacher evaluation policies on student mathematics and reading achievement. Before analyzing the formal ITS and CITS models, I plotted the time trends in student achievement (Figures 2 through 5).

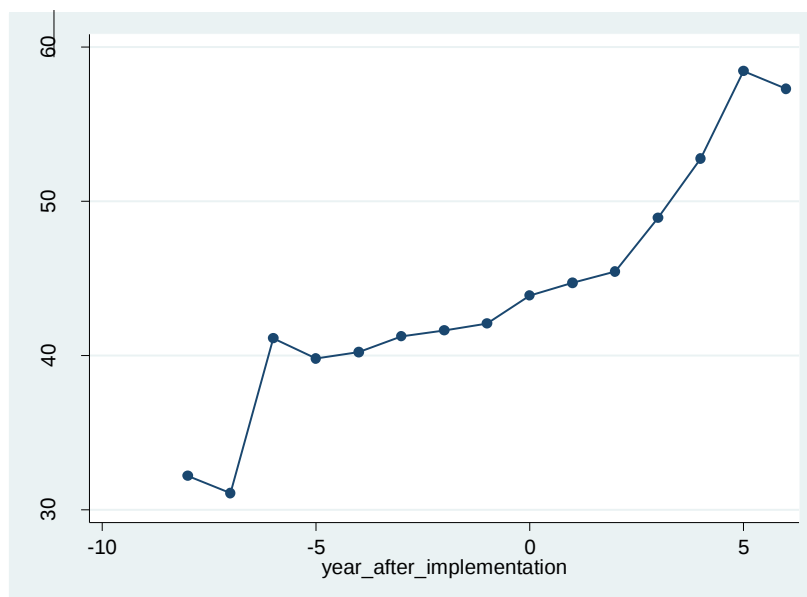


Figure 2. Time Trend in the Proportion of Proficient Students in Mathematics

Note. The vertical line is the year of implementation. Districts had different numbers of data points before and after policy implementation because they enacted the policies in different school years.

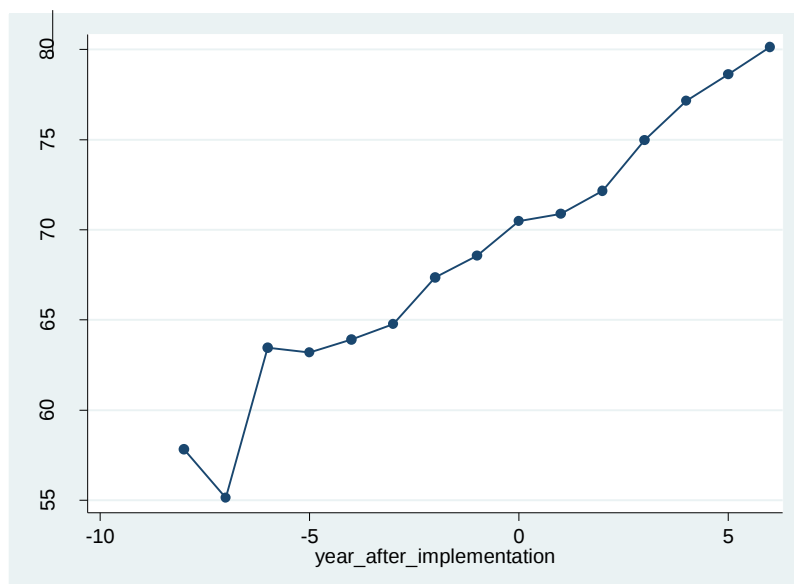


Figure 3. Time Trend in the Proportion of Proficient Students in Reading

Note. The vertical line is the year of implementation. Districts had different numbers of data points before and after policy implementation because they enacted the policies in different school years.

Figure 2 is based on the trend in the proportion of students who were proficient on the mathematics test at the district level before and after the policies, and Figure 3 is based on reading achievement. The trend in mathematics and reading achievement is roughly linear, which is an important condition for the ITS model. However, as Figures 4-5 (See Appendix D) show, the pre-treatment trend of the control group was not linear compared to that of treatment group.

Next, I proceed to the formal ITS model. As stated earlier, it should be noted that the results from this analysis are about the effects of teacher evaluation policies when they are fully implemented, compared to ones when they are partially implemented. Tables 13 and 14 report the results using the proportion of proficient students in mathematics and reading achievement for different categories (i.e., all students, male/female, not economically disadvantaged/economically disadvantaged, White/Black/Hispanic, and elementary/middle school) at the district level. This analysis includes only compliers. All models include district fixed effects and standard errors are clustered at the district level. Using mean scaled scores did not lead to any significant changes in the results. The main interest here is the coefficients of Policy and Year-since-policy. The former estimates the changes before and after the policies, and the latter estimates the changes in the slope before and after the policies.

Table 13. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement
(Mathematics)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	All students	Male	Female	Not Economically Disadvantaged	Economically Disadvantaged
Standard Deviation	1.359*** (0.106)	1.257*** (0.0977)	1.278*** (0.130)	1.188*** (0.176)	1.261*** (0.125)
Year	1.425*** (0.234)	1.233*** (0.217)	1.612*** (0.267)	1.752*** (0.309)	1.428*** (0.246)
Year Since Policy	-0.594 (0.391)	-0.590 (0.381)	-0.531 (0.421)	-0.724 (0.439)	-0.647 (0.374)
Policy	-0.765 (0.670)	-0.608 (0.664)	-0.693 (0.738)	-0.589 (0.801)	-0.721 (0.635)
Total Enrollment	0.000663 (0.000811)	0.000271 (0.000769)	0.00111 (0.000894)	0.000945 (0.00112)	0.00105 (0.000695)
Proportion of FRL eligible Students	3.130 (4.393)	1.285 (4.610)	3.663 (4.810)	19.27* (8.645)	14.74*** (4.163)
Proportion of White Students	33.47*** (9.355)	36.93*** (8.665)	32.18** (10.24)	52.56*** (15.01)	32.92** (10.07)
N	650	650	650	648	650
	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
	White	Black	Hispanic	Elementary	Middle
Standard Deviation	1.417*** (0.112)	0.419* (0.158)	1.038*** (0.200)	1.816*** (0.146)	0.930*** (0.115)
Year	1.531*** (0.254)	1.189** (0.413)	1.560** (0.503)	1.669*** (0.246)	1.130*** (0.310)
Year Since Policy	-0.656 (0.393)	-0.134 (0.898)	-1.919*** (0.525)	-0.920 (0.474)	-0.184 (0.438)
Policy	-0.704 (0.654)	-0.459 (1.700)	0.907 (0.939)	-0.677 (0.820)	-0.849 (0.811)
Total Enrollment	0.000464 (0.000852)	0.00145 (0.00115)	0.000564 (0.00135)	0.000616 (0.000985)	0.000634 (0.000854)
Proportion of FRL eligible Students	3.695 (4.845)	-0.505 (7.200)	-3.454 (11.69)	-2.109 (5.602)	6.747 (4.727)
Proportion of White Students	31.73** (11.59)	20.93* (10.21)	-0.321 (26.50)	27.47* (13.32)	38.78*** (9.859)
N	650	407	493	650	650

Table 14. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Reading)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	All students	Male	Female	Not Economically Disadvantaged	Economically Disadvantaged
Standard Deviation	-0.436*** (0.0809)	-0.229* (0.0908)	-0.636*** (0.0858)	-0.334 (0.183)	-0.423*** (0.102)
Year	1.919*** (0.161)	1.899*** (0.153)	2.008*** (0.196)	1.933*** (0.268)	2.594*** (0.189)
Year Since Policy	0.0430 (0.250)	0.302 (0.296)	-0.197 (0.248)	-0.0915 (0.303)	-0.200 (0.269)
Policy	0.266 (0.422)	0.145 (0.544)	0.269 (0.437)	-0.182 (0.517)	0.421 (0.508)
Total Enrollment	0.000466 (0.000561)	0.000344 (0.000583)	0.000555 (0.000588)	0.00113 (0.000704)	0.00114 (0.000607)
Proportion of FRL eligible Students	-5.954* (2.911)	-3.960 (3.374)	-7.520* (3.382)	14.54** (4.765)	2.794 (4.117)
Proportion of White Students	28.61** (9.002)	35.59*** (7.380)	24.58+ (13.52)	37.95** (13.08)	29.47** (9.034)
N	650	650	650	648	650
	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
	White	Black	Hispanic	Elementary	Middle
Standard Deviation	-0.486*** (0.0855)	0.0422 (0.0680)	0.0911 (0.215)	0.0495 (0.102)	-0.199** (0.0689)
Year	1.769*** (0.167)	2.328*** (0.534)	3.401*** (0.545)	1.742*** (0.175)	2.359*** (0.197)
Year Since Policy	0.0715 (0.270)	-0.208 (0.610)	-1.359 (0.822)	-0.526* (0.235)	0.403 (0.309)
Policy	0.390 (0.452)	-0.558 (1.028)	1.780 (1.590)	-0.0242 (0.421)	0.287 (0.560)
Total Enrollment	0.000282 (0.000611)	0.00197* (0.000923)	0.000445 (0.00144)	0.000305 (0.000718)	0.000721 (0.000589)
Proportion of FRL eligible Students	-6.296 (3.318)	2.919 (6.053)	-0.890 (12.21)	1.395 (3.386)	-5.080 (3.371)
Proportion of White Students	14.40 (7.965)	23.27* (9.962)	43.18 (33.91)	30.88** (9.144)	30.48** (9.813)
N	650	407	491	650	650

Note. The outcome is the proportion of proficient students at the district level from the 2007-08 school year to the 2013-14 school year. District fixed effects were included for all models and standard errors are clustered at the district level. Number of districts is 80-93 depends on the categories. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In terms of both mathematics and reading achievement, the implementation of teacher evaluation policies had no statistically significant impact on student achievement. The exceptions are the impact of the policies on mathematics achievement of Hispanic students and those on reading achievement of elementary school students. The coefficients of Year-since-policy were negative and significant for both cases. After the implementation of the policies, the growth in student achievement across years slowed down for these students. Based on the Robustness Indices (Frank, 2000; Frank et al., 2008; Frank et al., 2013), the effects on the Hispanic students seem to be quite strong; 46% of the sample would need to be replaced to invalidate the inference. Additionally, those pertaining to elementary school students are relatively weak; only 12% of the sample would need to be replaced to invalidate the inference. However, this analysis is using only a subset of the sample, and for students in general, the implementation of teacher evaluation policies did not have any impact on student achievement.

The results were quite different when I used the CITS model, instead of using ITS models [See Appendix E for the results]. The implementation of policies had a negative and statistically significant impact; the policies decreased the proportion of proficient students by 13.11% at the district level ($p < 0.05$). However, the coefficient of the slope change is also statistically significant, but positive ($b = 2.775$, $p < 0.05$). That is, students in the treatment group scored lower than their counterparts in the control group when the district first implemented the policies, controlling for their achievement trends before the policies were enacted as well as other district characteristics. However, their achievement rapidly increased after the policies were implemented, compared to their achievement trends before implementation and the control group students' trends. This same pattern was found for female, economically disadvantaged, White, Black, and elementary school students. Although the patterns in the effects of the policies remained similar among those groups, the effects of the policies were more salient for female and Black students. The policies did not affect the mathematics achievement of male, not economically disadvantaged, Hispanic, or middle school students. In contrast, the effects of the policies on reading achievement were not statistically significant for any groups.

One of the main caveats for this analysis is that the cut-off year for the control group (i.e., 2011-12) is arbitrary; there is no single time point when all districts implemented the policies as the law required. Accordingly, as a robustness check, I used the 2012-13 school year as the cut-off year for the control group and ran the same models above. Almost all main coefficients were no longer significant, but the directions of the coefficients stayed the same [See Appendix F for the details]. For mathematics achievement, the policies had a negative impact, but the coefficients for the slopes were positive. The results about reading achievement were similar as well. As noted above, however, as the main assumption of CITS model was not met, I consider the ITS results to be the main results for the current study.

Discussion

Teacher evaluation policies are one of the most controversial issues in the past few decades in that they accelerate the movement toward tight coupling that has been pursued by different types of standards-based reforms, such as *NCLB*. As envisioned in classic studies on school organizations, classrooms might be the hardest place to change with new interventions (Coburn, 2004). In this sense, investigating teacher evaluation policies, which deal with the very core of schooling, involves more than just examining a single policy. It helps us understand how school organizations, which have long been regarded as loosely coupled systems, respond to a force aiming towards tightly coupled systems. The current study examined this question by investigating three research questions in Michigan school districts: 1) Were there clear variations in the implementation of teacher evaluation policies? 2) Which factors potentially affected such decisions? and 3) How does implementation of teacher evaluation policies affect student achievement?

Combining survey data from 101 district administrators with multiple data sets, such as fiscal resources available in each district and student MEAP test scores, this study generated three main findings. First, there were clear variations in the implementation of teacher evaluation policies across districts, in terms of the timing of the policies and whether districts used teacher evaluation ratings for making decisions about teacher dismissal. Some school districts implemented the policies even before it was mandatory to do so; others still had not enacted the policies by 2015-16 in the way required by the state.

Such variations in teacher evaluation policies are also documented in studies at the state level (Gagnon, Hall, & Marion, 2016; Steinberg & Donaldson, 2016). This study shows that, even within the same state and with the same regulations, districts implemented the policies in various ways. That is, the system of Michigan school districts and the state is still loosely coupled, in terms of teacher evaluation policies, despite the strong pressure for tighter coupling embedded in the policies themselves.

This is attributable to two factors. First, Michigan school districts have historically been under local control (Spillane, 1996). Districts' histories have been regarded as an important factor for their response to *NCLB* (Terry, 2010) and their implementation of an innovation (Anderson-Butcher et al., 2010). Since school districts in Michigan have been heavily influenced by local control, rather than state-level control, they might continue to separate themselves from state-level regulations, focusing more on their own needs. Second, the law itself opens the possibilities of variation at the district level. The law did not designate certain tools for classroom observation or student growth measures. Without unified tools, districts had to find their own tools and train their evaluators accordingly. More importantly, there was no systemic monitoring or support for this process from the state government. These aspects of teacher evaluation jointly keep the system loosely coupled, in terms of teacher evaluation policies. In this sense, it is worthwhile to point out Elmore's (2000) argument about the buffering effects of loose coupling;

(b)uffering consists of creating structures and procedures around the technical core of teaching that, at the same time, (1) protect teachers from outside intrusions in their highly uncertain and murky work, and (2) create the appearance of rational management of the technical core, so as to allay the uncertainties of the public about the actual quality or legitimacy of what is happening in the technical core (p. 6).

It is plausible that the appearance of implementing teacher evaluation provides the buffering effects for some school districts, since it produces an image of rational management of teaching and being controlled by the state government for the public. In fact, what districts do under the name of teacher evaluation might not be the same across districts and may be different from what the state policy makers intended.

The second part of this study addresses the question about which districts were more or less likely to need such buffering effects. Districts with more White students and fiscal resources were more likely to be early adopters of the policies, controlling for student achievement, administrators' state-level committee membership, and other district demographics. These districts implemented the policies even before the state enacted the policies; that is, they were active in making anticipatory changes in terms of teacher evaluation policies. This finding is consistent with Donaldson and colleagues (2016); teachers' learning opportunity varies at the district level, and districts serving a greater share of students of color provided teachers fewer and lower-quality opportunities to learn based on teacher evaluation results. There are two plausible explanations for this. First, given that White parents tend to be more actively involved in schools (Catsambis & Garland, 1997), districts serving more White students may have received more pressure to implement the teacher evaluation policies. Second, given the resources that each district needed to spend on implementing the policies, districts with a large amount of resources available could implement the policies earlier without critical reallocation of the resources for other initiatives. This finding also suggests that, without enough pressures and resources, districts would not be motivated to actively implement teacher evaluation policies.

Analyses of late adopters and non-compliers show which districts might have been passive in response to state-level initiatives. They implemented the policies later than the state required them to do or did not implement them as they were required. For these districts, the state system was indeed loosely coupled. Districts serving more low-achieving students were more likely to be non-compliers. This result indicates an important pattern in the implementation of teacher evaluation policies at the district level. Districts facing strong pressure to raise student achievement might lack the capacity or motivation to implement the policies, although the goal of the policies targets the issue (i.e., student performance) with which they are struggling. As noted in the results section, between capacity and motivation, a potential lack of motivation was revealed during the interviews. However, it is also possible that these districts did not have sufficient capacity to implement the policies, and such lack of capacity was manifested by low achievement. As Cawelti and Protheroe (2007) argued, school districts that are successful in school

improvement tend to have “strong leadership by a superintendent and school board willing and able to publicly recognize challenges, develop a plan for reform, and build support for needed changes” (p. 29). That is, in those districts in my study, the leadership was not strong enough to implement the policies on time, due to the challenges associated with the policies, such as resources, teachers’ resistance, and time. Although I included the membership on a state-level committee (MASA) as a proxy for some aspects of leadership, it might not be enough to capture the nuanced aspects of leadership at the district level.

Lastly, I analyzed the effects of teacher evaluation policies on student achievement in mathematics and reading. Since the intervention itself and fidelity to the policies were clearly different across districts, this analysis is an exploratory and intent-to-treat analysis. The results found that the implementation of teacher evaluation policies had no significant effects on student achievement. This is quite surprising given the results of the previous studies about the positive effects of teacher evaluation (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). There are two ways to explain such a difference. First, the focus in the previous studies was at the individual teacher level or school level. If a district enacted the policies for all schools, not just a group of teachers or a few schools, the district might need to reallocate their resources (i.e., time and budget) and create a system for implementing teacher evaluation policies. This might be a significant shift, and districts would need to adjust themselves accordingly. This can account for the weak effects of the policies; districts need to endure the transition costs for implementing the policies. Second, the control groups in the previous studies were totally free from teacher evaluation, while this study is based on a comparison between partial implementation and full implementation of teacher evaluation policies. It is possible that the policies had a significant effect on student achievement when districts first enacted one or two components of them. After enacting some aspects of the policies, fully implementing the policies might not produce any new effects.

There are a few limitations of this study that are worthwhile to note. First, the main instrument of the study, the survey of district administrators, asked respondents to think back a few years ago, which may reduce the accuracy of the data. However, there was no other available data containing detailed information about the timing of implementation. Although districts were required to report teacher

evaluation ratings starting with the 2011-12 school year, it is possible that they did not evaluate all teachers or use teacher evaluation results at all. In this situation, asking the people who worked in each district might be the best way to collect this type of data. Based on the survey, the average years working in the same district of the respondents was 12.05 years, which covered most of the critical years of the enactment of teacher evaluation policies.

Second, there are only seven years of student achievement data available for this analysis, and districts had different numbers of time points depending on when they enacted the policies. For some districts that implemented the policies too early or too late, there were not enough time points to calculate the slope before or after the policies. For example, if a district implemented the policies in 2009-10, there were only two years of data before the policies (2007-08 and 2008-09); thus, the trend before the policies might be less reliable in such a case. This was inevitable in that the student achievement data is only available from 2007-08 to 2013-14, and M-STEP data, started in 2014-15, is not compatible with MEAP test scores. The different time points at which districts enacted the policies also made it hard to set a cut-off point for control group districts for the CITS models.

Third, the definition of the implementation of the policies in this study was narrow. There are so many different ways to define the implementation of teacher evaluation policies; such as which evaluation tools were used, whether student test scores were included, and how professional development programs were devised based on the evaluation results. It is possible that some districts were categorized as non-compliers, but in fact implemented the policies in a more effective way to improve student achievement. I use the definition derived from the state law, because of the framework of loose coupling. My focus is not whether the policies were implemented in a way that they indeed improved teaching quality; my focus was whether the policies were implemented as required. However, I acknowledge that asking the former question might be also very important for the enhancing the effects of teacher evaluation policies.

Given these limitations, I suggest three directions for the future research. First, in order to understand the long-term effects of the policies, an analysis using more years of student achievement data would be fruitful. This question also implies that teacher evaluation policies are not short-term policies,

which can be quickly implemented and evaluated in a short time. This idea can be applied to other educational reforms as well; “Policy Churn” fails to result in steady school improvement (Hess, 1998; Newmann, King, & Youngs, 2000). As in teacher evaluation settings, institutional changes require time for districts and schools to adjust themselves.

Second, the implementation of teacher evaluation policies in different districts deserves more in-depth study. Some districts might have been able to maximize the positive effects of the policies, and it is important to examine their strategies to boost the effects of the policies. To be specific, researchers should investigate how these districts effectively reconcile the needs of their own community with the state-level requirement, which types of systemic changes they made in relation to teacher evaluation policies, and how those changes contribute to student achievement. A case study approach using mixed methods would be appropriate for this study.

Third, the analysis of the effects of teacher evaluation at the school and teacher levels might add more nuance to the current study. Principals at each school may have implemented the policies differently than districts did. Teacher evaluation at some schools may produce a positive effect on student learning as it is treated as an opportunity for professional learning. This applies to the individual teacher level. For example, teacher evaluation may be more effective for early career teachers, who would be more reactive to the evaluation compared to experienced teachers. In the same vein, teacher evaluation may have a stronger impact on teachers who teach tested grades and subjects.

As noted earlier, there is a high probability that teacher evaluation policies will become a subject of local control for many jurisdictions under the new ESSA. This study has documented possible variation in the implementation of teacher evaluation policies and the effects of the policies on student achievement in a state where districts maintain a high level of local control setting. In terms of teacher evaluation policies, districts in Michigan were still loosely coupled even after the new teacher evaluation policies were launched and various factors might have affected their decisions regarding enactment of the policies. Whether this is ultimately conducive or detrimental for teachers and students is beyond the scope of this study, but the results of this study indicates that implementing the policies as required (i.e.,

enacting a tightly coupled policy) did not have any effects on student achievement, despite a considerable amount of resources that the policies required of districts.

NOTES

NOTES

¹ Given that there was an exceptional clause in the law that districts that had ongoing Collective Bargaining Agreement (CBA) in the 2011-12 school year could postpone the implementation to the 2012-13 school year, districts that implemented the policies in 2012-13 were not coded as late adopters. However, there was no clear pattern between the timing of the implementation and districts' CBA.

² The main reason for using the 2007-08 achievement data is the fact that later years' student achievement scores might be, in part, the results of the policies. In addition, since student achievement at the district level is relatively stable across years, I treated this variable as a district characteristic.

APPENDICES

Appendix A

Mich. Comp. Laws § 380.1249

- (a) Evaluates the teacher's or school administrator's job performance at least annually while providing timely and constructive feedback.
- (b) Establishes clear approaches to measuring student growth and provides teachers and school administrators with relevant data on student growth.
- (c) Evaluates a teacher's or school administrator's job performance, using multiple rating categories that take into account student growth and assessment data. Student growth must be measured using multiple measures that may include student learning objectives, achievement of individualized education program goals, nationally normed or locally developed assessments that are aligned to state standards, research-based growth measures, or alternative assessments that are rigorous and comparable across schools within the school district, intermediate school district, or public school academy. If the performance evaluation system implemented by a school district, intermediate school district, or public school academy under this section does not already include the rating of teachers as highly effective, effective, minimally effective, and ineffective, then the school district, intermediate school district, or public school academy shall revise the performance evaluation system not later than September 19, 2011 to ensure that it rates teachers as highly effective, effective, minimally effective, or ineffective.
- (d) Uses the evaluations, at a minimum, to inform decisions regarding all of the following:
 - (i) The effectiveness of teachers and school administrators, ensuring that they are given ample opportunities for improvement.
 - (ii) Promotion, retention, and development of teachers and school administrators, including providing relevant coaching, instruction support, or professional development.
 - (iii) Whether to grant tenure or full certification, or both, to teachers and school administrators using rigorous standards and streamlined, transparent, and fair procedures.

(iv) Removing ineffective tenured and untenured teachers and school administrators after they have had ample opportunities to improve, and ensuring that these decisions are made using rigorous standards and streamlined, transparent, and fair procedures.

Appendix B

Interview Protocol for District Administrators

1. How long have you worked in this district? What is your main role?
2. Could you describe the current teacher evaluation policy for your school district? (i.e., components, weight for each component, differentiation between tenured vs. probationary teachers and between teachers in tested/non-tested subjects)
3. Which observation rubric are you using? How did you choose the rubric?
4. Could you describe how your district uses the results from teacher evaluations?
5. In which school year did your district start to evaluate all teachers every year and make-decisions about teacher compensation, assignment, dismissal, based on teacher evaluation result for the first time?
6. Could you describe the decision making process in your district regarding teacher evaluation policy (rubric, timing, weight, use of results....) What are the reasons why your district implemented the policy at that time point? (e.g., NCLB, state mandate, superintendent background, concern about teacher tenure, school board member election, student achievement scores, budget, etc.)
7. Could you describe how teacher evaluation policy affects teachers in your district?
8. In your district, which individuals were responsible for deciding to make this change and/or implementing the new teacher evaluation policy? (e.g., superintendent, school board members, principals, teacher union leaders)
9. Could you describe any policy changes or other events that might have affected teachers' instruction and/or students' MEAP test scores since the 2010-11 school year?
10. Was there any training for evaluators in your district? If so, could you describe the training process?
11. Was there any teacher involvement in decision making process regarding teacher evaluation policy in your district?

12. How are students' MEAP test scores used in your district?
13. From your perspectives, would the implementation of the teacher evaluation policy at district level increase students' test scores?

Appendix C

Multicollinearity Check

Table 15. Multicollinearity Check for Logistic Regressions

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
Proportion of proficient students (%)	1.027 (0.0434)	0.984 (0.0533)	1.002 (0.0699)	0.962 (0.0806)	0.958 (0.0779)
Standard deviation	0.918 (0.228)	1.119 (0.304)	1.058 (0.314)	1.191 (0.44)	1.228 (0.460)
Total enrollment	1 (0.000188)	1 (0.000156)	1 (0.000177)	0.999 (0.000469)	0.999 (0.000437)
Proportion of White students (%)		1.064* (0.0268)	1.066* (0.0266)	1.118* (0.045)	1.118* (0.046)
Proportion of free and reduced lunch eligible students (%)			1.015 (0.0429)	1.021 (0.053)	1.028 (0.0558)
District total revenue in 2007-08 (log)				16.31* (23.08)	18.86* (25.51)
Suburban					1.355 (1.05)
Rural					2.121 (1.943)

Note. The outcome for these models is districts being early adopters of the policies. Cluster robust errors at the county level were used. The student achievement variables used in these models are based on MEAP test scores in mathematics during the 2007-08 school year. The coefficients are expressed in odds-ratios for ease of interpretation. Model 5 is the final model. N=95. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix D

Time Trend Using CITS Model

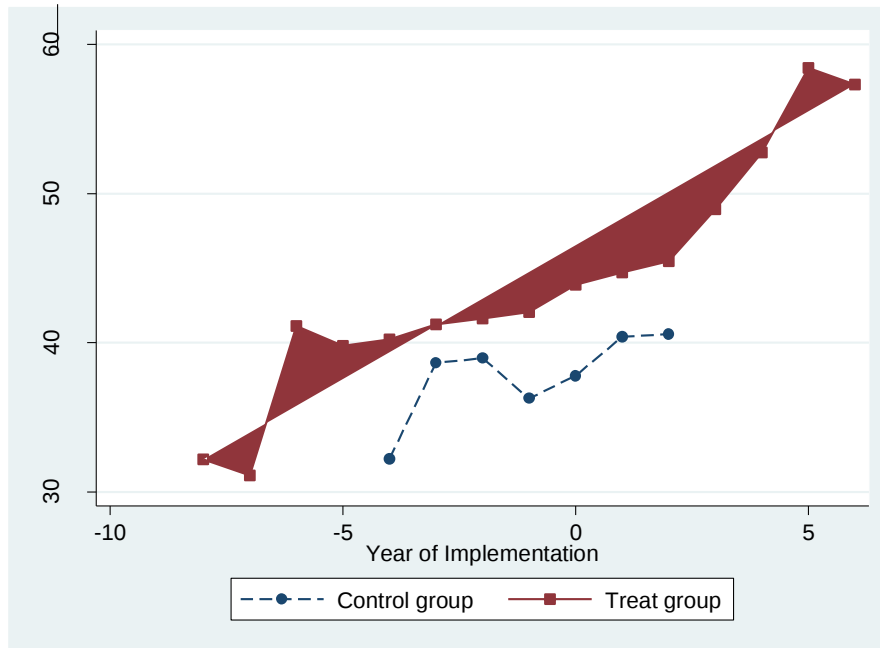


Figure 4. Time Trend in the Proportion of Proficient Students in Mathematics (CITS model)

Note. The vertical line is the year of implementation. The treatment group had more time points because their years of implementation varied across districts while those of the control group were fixed as the 2011-12 school year. Thus, there were only four time points before and three time points after policy implementation for all control group districts, while treatment group districts had varying numbers of points before and after policy implementation.

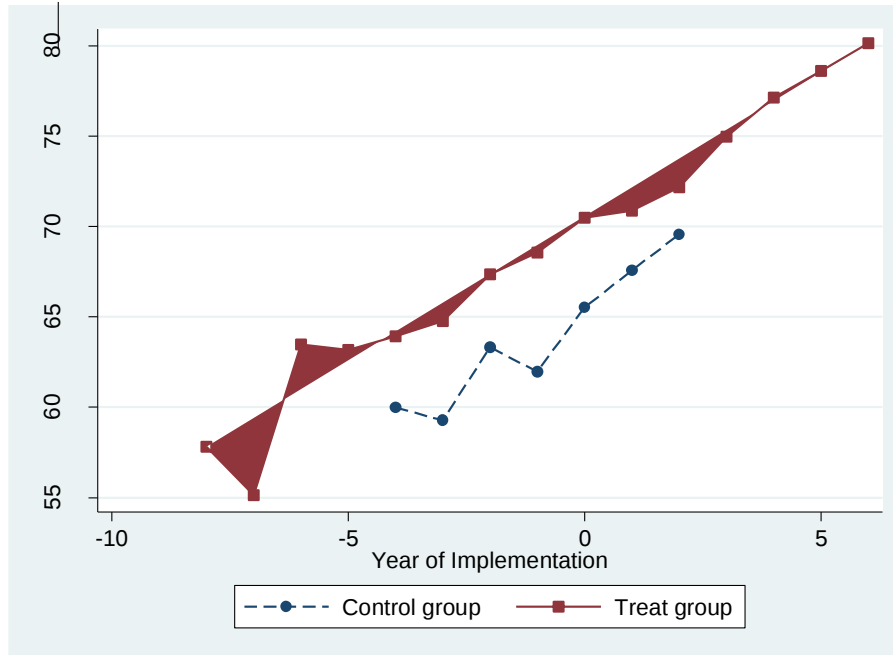


Figure 5. Time Trend in the Proportion of Proficient Students in Reading (CITS model)

Note. The vertical line is the year of implementation. The treatment group had more time points because their years of implementation varied across districts while those of the control group were fixed as the 2011-12 school year. Thus, there were only four time points and three time points after policy implementation for all control group districts, while treatment group districts have different number of points before and after the policies.

Appendix E

Results using Comparative Interrupted Time Series Models

Table 16. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement
(Mathematics)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	All students	Male	Female	Not Economically Disadvantaged	Economically Disadvantaged
Standard Deviation	1.477*** (0.114)	1.354*** (0.104)	1.391*** (0.137)	1.304*** (0.185)	1.345*** (0.133)
Year	2.584*** (0.537)	2.242*** (0.479)	2.89*** (0.672)	2.357*** (0.631)	2.937*** (0.793)
Post Policy	16.34** (6.137)	11.56 (6.909)	19.26** (6.01)	13.49 (7.108)	16.24** (5.489)
Year*T	-1.048* (0.523)	-0.937* (0.466)	-1.162 (0.658)	-0.564 (0.585)	-1.452 (0.78)
Year*Post Policy	-3.659** (1.331)	-2.837* (1.427)	-4.132** (1.353)	-2.936 (1.543)	-3.931 (1.279)
T*Post Policy	-13.11* (6.252)	-8.684 (6.998)	-16.17* (6.22)	-10.17 (7.253)	-13.38* (5.518)
Year*T*Post Policy	2.775* (1.333)	2.507 (1.429)	3.298* (1.364)	2.05 (1.533)	3.124* (1.267)
Total Enrollment	0.000768 (0.000788)	0.000371 (0.000745)	0.00121 (0.000873)	0.00099 (0.00108)	0.00111 (0.000681)
Proportion of FRL eligible Students	1.948 (4.103)	0.762 (4.32)	2.248 (4.509)	19.15* (8.193)	13.69*** (3.994)
Proportion of White Students	32.40*** (8.464)	36.26*** (7.977)	30.71** (9.368)	50.89*** (14.19)	30.98** (9.317)
N	706	706	706	704	706
	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
	White	Black	Hispanic	Elementary	Middle
Standard Deviation	1.540*** (0.118)	0.353** (0.125)	1.137*** (0.204)	1.988*** (0.153)	0.969*** (0.113)
Year	2.726*** (0.594)	3.623* (1.692)	2.574 (1.646)	2.362*** (0.444)	2.481** (0.866)
Post Policy	17.13** (6.165)	15.48** (4.655)	9.206 (15.08)	18.28*** (4.747)	11.46 (9.641)
Year*T	-1.072 (0.579)	-2.606 (1.724)	-1.162 (1.612)	-0.599 (0.443)	-1.246 (0.851)
Year*Post Policy	-3.871** (1.37)	-4.294* (1.831)	-3.133 (3.907)	-3.943*** (1.083)	-2.744 (2.014)
T*Post Policy	-13.4* (6.256)	-16.9* (7.15)	-1.226 (15.2)	-13.62** (5.111)	-10.2 (9.76)
Year*T*Post Policy	2.89* (1.369)	4.47* (2.021)	1.472 (3.888)	2.742* (1.119)	2.302 (2.017)
Total Enrollment	0.000559 (0.000831)	0.00149 (0.00115)	0.0005 (0.00132)	0.000577 (0.00096)	0.000862 (0.000842)

Table 16 (cont'd)

	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
Proportion of FRL eligible Students	2.078 (4.51)	4.026 (7.749)	-2.232 (11.82)	-3.287 (5.433)	5.338 (4.348)
Proportion of White Students	30.01** (10.55)	23.28* (10.6)	-6.132 (26.62)	25.78* (12.34)	38.05*** (9.182)
N	706	435	524	706	706

Table 17. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Reading)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	All students	Male	Female	Not Economically Disadvantaged	Economically Disadvantaged
Standard Deviation	-0.472*** (0.0834)	-0.312** (0.0995)	-0.605*** (0.089)	-0.363 (0.192)	-0.406*** (0.103)
Year	1.02*** (0.255)	0.601 (0.309)	1.555*** (0.259)	1.085*** (0.319)	1.739*** (0.477)
Post Policy	-5.787 (3.91)	-11.71* (5.781)	0.646 (3.041)	-5.195 (3.997)	-3.034 (4.949)
Year*T	0.799** (0.251)	1.143*** (0.33)	0.409 (0.225)	0.726* (0.293)	0.727 (0.482)
Year*Post Policy	1.565* (0.766)	2.65* (1.077)	0.322 (0.601)	1.311* (0.633)	0.837 (0.991)
T*Post Policy	4.785 (4.154)	8.868 (6.113)	0.000903 (3.23)	4.169 (4.047)	2.946 (5.227)
Year*T*Post Policy	-1.308 (0.803)	-2.016 (1.136)	-0.431 (0.624)	-1.159 (0.628)	-0.775 (1.033)
Total Enrollment	0.000473 (0.000552)	0.000316 (0.000564)	0.00061 (0.000588)	0.0011 (0.000667)	0.00103 (0.000607)
Proportion of FRL eligible Students	-4.158 (2.758)	-1.957 (3.199)	-5.733 (3.233)	16.64*** (4.617)	5.194 (3.94)
Proportion of White Students	28.8** (8.935)	35.7*** (7.358)	25.05 (12.89)	37.27** (12.87)	28.14** (9.546)
N	706	706	706	704	706
	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
	White	Black	Hispanic	Elementary	Middle
Standard Deviation	-0.519*** (0.0865)	0.0409 (0.0564)	0.0709 (0.209)	0.0718 (0.0984)	-0.246*** (0.0706)
Year	0.955*** (0.275)	2.039 (2.583)	-2.391 (4.279)	1.084*** (0.294)	1.305*** (0.305)
Post Policy	-6.555 (4.143)	19.39 (10.27)	10.07 (10.77)	2.636 (2.831)	-11.61 (6.897)
Year*T	0.726** (0.276)	0.242 (2.607)	5.805 (4.311)	0.694* (0.293)	0.899** (0.307)
Year*Post Policy	1.635* (0.819)	-3.008 (3.802)	0.445 (3.607)	-0.197 (0.612)	2.778* (1.248)
T*Post Policy	5.669 (4.405)	-19.6 (40.64)	-1.552 (12.11)	-0.115 (3.06)	8.341 (7.088)
Year*T*Post Policy	-1.372 (0.86)	2.911 (3.833)	-1.931 (3.752)	-0.382 (0.641)	-2.018 (1.277)
Total Enrollment	0.000321 (0.000596)	0.00215* (0.000956)	0.000402 (0.0015)	0.000374 (0.000721)	0.000705 (0.000579)
Proportion of FRL eligible Students	-4.469 (3.15)	5.464 (6.648)	-5.721 (11.3)	0.428 (3.109)	-1.218 (3.612)
Proportion of White Students	15.08 (7.826)	25.06* (10.54)	33.32 (34.12)	29.3*** (8.428)	31.66*** (9.926)
N	706	435	522	706	706

Table 17 (cont'd)

Note. The outcome is the proportion of proficient students at the district level from the 2007-08 school year to the 2013-14 school year. District fixed effects were included for all models and standard errors are clustered at the district level. Number of districts=101. * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Appendix F

Results using CITS Model with 2012-13 School Year as Cut-Off Point

Table 18. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement
(Mathematics)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	All students	Male	Female	Not Economically Disadvantaged	Economically Disadvantaged
Standard Deviation	1.484*** (0.116)	1.368*** (0.106)	1.39*** (0.138)	1.303*** (0.186)	1.355*** (0.135)
Year	2.244*** (0.396)	1.812*** (0.385)	2.592*** (0.469)	2.068*** (0.482)	2.326*** (0.498)
Post Policy	10.05 (10.01)	1.139 (10.8)	19.64* (9.896)	17.52 (9.722)	6.941 (9.639)
Year*T	-0.714 (0.397)	-0.515 (0.386)	-0.871 (0.474)	-0.279 (0.449)	-0.852 (0.492)
Year*Post Policy	-2.484 (1.739)	-0.982 (1.899)	-3.984* (1.707)	-3.34 (1.752)	-2.118 (1.647)
T*Post Policy	-6.817 (10.22)	1.762 (10.99)	-16.58 (10.14)	-14.22 (9.945)	-4.097 (9.784)
Year*T*Post Policy	1.599 (1.766)	0.197 (1.923)	3.156 (1.742)	2.459 (1.772)	1.314 (1.664)
Total Enrollment	0.000771 (0.000787)	0.000375 (0.000743)	0.00121 (0.000873)	0.000986 (0.00108)	0.00111 (0.00068)
Proportion of FRL eligible Students	2.321 (4.102)	1.291 (4.316)	2.498 (4.502)	19.38* (8.179)	14.32*** (4.036)
Proportion of White Students	32.43*** (8.503)	36.25*** (8.013)	30.86** (9.4)	51.1*** (14.17)	31.1** (9.372)
N	706	706	706	704	706
	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
	White	Black	Hispanic	Elementary	Middle
Standard Deviation	1.545*** (0.12)	0.345** (0.124)	1.117*** (0.202)	1.986*** (0.155)	0.979*** (0.114)
Year	2.318*** (0.407)	2.372* (1.173)	0.837 (0.636)	2.026*** (0.353)	2.182*** (0.61)
Post Policy	12.04 (9.992)	7.076 (21.77)	24.41 (25.13)	21.74 (13.11)	-4.402 (14.67)
Year*T	-0.672 (0.404)	-1.382 (1.152)	0.535 (0.596)	-0.269 (0.359)	-0.953 (0.615)
Year*Post Policy	-2.828 (1.756)	-2.166 (2.874)	-4.267 (4.484)	-4.23 (1.991)	-0.162 (2.611)
T*Post Policy	-8.318 (10.18)	-8.632 (22.13)	-16.65 (25.42)	-17.11 (13.39)	5.685 (14.83)
Year*T*Post Policy	1.849 (1.779)	2.37 (2.932)	2.659 (4.541)	3.035 (2.039)	-0.284 (2.631)
Total Enrollment	0.000561 (0.000831)	0.00149 (0.00115)	0.000489 (0.00132)	0.000573 (0.000959)	0.000871 (0.000843)

Table 18 (cont'd)

	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
Proportion of FRL eligible Students	2.49 (4.502)	4.764 (8.341)	-0.173 (11.5)	-3.018 (5.434)	5.753 (4.323)
Proportion of White Students	30.12** (10.59)	23.34* (10.56)	-4.088 (26.36)	26.01* (12.34)	37.89*** (9.226)
N	706	435	524	706	706

Table 19. The Effect of Implementation of Teacher Evaluation Policies on Student Achievement (Reading)

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	All students	Male	Female	Not Economically Disadvantaged	Economically Disadvantaged
Standard Deviation	-0.471*** (0.0825)	-0.319** (0.0974)	-0.6*** (0.0906)	-0.364 (0.193)	-0.406*** (0.103)
Year	1.454*** (0.237)	0.961*** (0.265)	2.006*** (0.251)	1.389*** (0.303)	1.982*** (0.396)
Post Policy	-6.717 (9.123)	-15.89 (11.99)	2.695 (9.588)	-7.223 (8.53)	-3.277 (10.37)
Year*T	0.373 (0.24)	0.786** (0.294)	-0.0325 (0.225)	0.427 (0.288)	0.489 (0.412)
Year*Post Policy	1.404 (1.471)	3.028 (1.959)	-0.297 (1.509)	1.405 (1.376)	0.706 (1.679)
T*Post Policy	5.75 (9.224)	13.06 (12.14)	-1.993 (9.636)	6.216 (8.379)	3.207 (10.52)
Year*T*Post Policy	-1.156 (1.492)	-2.395 (1.996)	0.177 (1.517)	-1.258 (1.333)	-0.648 (1.71)
Total Enrollment	0.000477 (0.000552)	0.000318 (0.000564)	0.000613 (0.000588)	0.0011 (0.000668)	0.00103 (0.000607)
Proportion of FRL eligible Students	-4.519 (2.716)	-2.335 (3.164)	-6.081 (3.186)	16.38*** (4.551)	4.984 (3.912)
Proportion of White Students	28.58*** (8.953)	35.38*** (7.406)	24.9 (12.88)	37.07** (12.91)	28.01** (9.566)
N	706	706	706	704	706
	Model (6)	Model (7)	Model (8)	Model (9)	Model (10)
	White	Black	Hispanic	Elementary	Middle
Standard Deviation	-0.521*** (0.086)	0.0441 (0.0597)	0.0631 (0.208)	0.0697 (0.1)	-0.257*** (0.0706)
Year	1.329*** (0.254)	3.302*** (0.876)	0.548 (2.231)	1.291*** (0.192)	1.913*** (0.343)
Post Policy	-10.14 (9.212)	-8.13 (32.93)	6.781 (21.79)	16.01 (9.428)	-24.83** (12.76)
Year*T	0.358 (0.264)	-1.007 (0.941)	2.933 (2.294)	0.494* (0.205)	0.298 (0.351)
Year*Post Policy	1.915 (1.506)	0.235 (4.249)	-1 (3.628)	-2.344 (1.48)	4.345* (2.11)
T*Post Policy	9.279 (9.312)	8.005 (32.96)	1.96 (22.53)	-13.47 (9.497)	21.55 (12.81)
Year*T*Post Policy	-1.657 (1.526)	-0.351 (4.267)	-0.54 (3.768)	1.762 (1.493)	-3.582 (2.118)
Total Enrollment	0.000325 (0.000597)	0.00216* (0.000957)	0.000415 (0.00151)	0.000366 (0.000721)	0.000711 (0.000579)
Proportion of FRL eligible Students	-4.796 (3.108)	4.79 (6.536)	-9.515 (12.09)	0.178 (3.093)	-1.831 (3.471)
Proportion of White Students	14.82 (7.849)	24.29* (10.68)	30.69 (34.34)	29.46*** (8.371)	30.95** (10.03)
N	706	435	522	706	706

Table 19 (cont'd)

Note. The outcome is the proportion of proficient students at the district level from the 2007-08 school year to the 2013-14 school year. District fixed effects were included for all models and standard errors are clustered at the district level. Number of districts=101. * $p<0.05$, ** $p<0.01$, *** $p<0.001$

REFERENCES

REFERENCES

- 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.
- Anderson-Butcher, D., Lawson, H. A., Iachini, A., Bean, G., Flaspohler, P. D., & Zullig, K. (2010). Capacity-related innovations resulting from the implementation of a community collaboration model for school improvement. *Journal of Educational and Psychological Consultation*, 20(4), 257-287.
- Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance some problems in the design and implementation of evaluation systems. *Educational Researcher*, 44(2), 77–86.
- Bidwell, C. E. (1965). The school as a formal organization. In March, J. G. (Ed.), *Handbook of organizations* (pp. 972–1019). Chicago, IL: Rand McNally.
- Boyd, W. L., & Wheaton, D. R. (1983). Conflict management in declining school districts. *Peabody Journal of Education*, 60(2), 25-36.
- Briggs, D., & Domingue, B. (2011). *A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center, University of Colorado.
- Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, 33(3), 378-398.
- Catsambis, S., Garland, J. (1997). *Parental involvement in students' education during middle school and high school: Report No. 181*. Baltimore, MD: Johns Hopkins University; Washington, DC: Howard University; Center for Research on the Education of Students Placed At Risk
- Cawelti, G. & Protheroe, N. (2007). The school board and central office in school improvement. In H. Walberg (Ed.), *Handbook on restructuring and substantial school improvement* (pp. 37–52). Lincoln, IL: Center on Innovation and Improvement.
- Cibulka, J. G. (1983). Explaining the problem: A comparison of closings in ten U.S. cities. *Education and Urban Society*, 15(2), 165-74.
- Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, 77(3), 211-244.
- Coburn, C. E., Touré, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *Teachers College Record*, 111(4), 1115-1161.
- Colby, S. A., Bradshaw, L. K., & Joyner, R. L. (2002). Teacher evaluation: A review of the literature. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA

- Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association of Supervision and Curriculum Development.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), 8-15.
- Darling-Hammond, L., Aneess, J., & Ort, S. W. (2002). Reinventing high school: Outcomes of the coalition campus schools project. *American Educational Research Journal*, 39(3), 639–673.
- Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and management*, 30(3), 418-446.
- Dembosky, J. W., Pane, J. F., Barney, H., & Christina, R. (2006). *Data driven decision making in Southwestern Pennsylvania school districts*. Santa Monica, CA: RAND.
- Donaldson, M. L., Woulfin, S. L., & Cobb, C. D. (2016). The structure and substance of teachers' opportunities to learn about teacher evaluation reform: promise or pitfall for equity? *Equity & Excellence in Education*, 49(2), 183-201.
- Dutro, E., Fisk, M. C., Koch, R., Roop, L. J., & Wixson, K. (2002). When state policies meet local district contexts: Standards-based professional development as a means to individual agency and collective ownership. *Teachers College Record*, 104(4), 787-811.
- Elmore, R. (2000). *Building a New Structure for School Leadership*. Washington, DC: Albert Shanker Institute.
- Firestone, W. A. (1985). The study of loose coupling: Problems, progress, and prospects. In A. Kerckhoff (Ed.), *Re-search in sociology of education and socialization* (Vol. 5, pp. 3-30). Greenwich, CT: JAI Press.
- Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research*, 29(2), 147-194.
- Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, 35(4), 437-460.
- Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? *Education Evaluation and Policy Analysis*, 30(1), 3–30.
- Fusarelli, L. D. (2002). Tightly coupled policy in loosely coupled systems: Institutional capacity and organizational change. *Journal of Educational Administration*, 40(6), 561-575.
- Gagnon, D. J., Hall, E. L., & Marion, S. (2016). Teacher evaluation and local control in the US: An investigation into the degree of local control afforded to districts in defining evaluation procedures for teachers in non-tested subjects and grades. *Assessment in Education: Principles, Policy & Practice*, Advance online publication. 1-17.
- Gamoran, A., & Dreeben, R. (1986). Coupling and control in educational organizations. *Administrative Science Quarterly*, 31(4). 612-632.

- Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, 44(2), 87–95.
- Goldspink, C. (2007). Rethinking educational reform a loosely coupled and complex systems perspective. *Educational Management Administration & Leadership*, 35(1), 27-50.
- Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, 26(1), 1-24.
- 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.
- Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, 116(1), 1-28.
- Hess, F. M. (1999). *Spinning wheels: The politics of urban school reform*. Washington, DC: Brookings Institution.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794–831.
- Honig, M. I. (2009). No small thing: School district central office bureaucracies and the implementation of new small autonomous schools initiatives. *American Educational Research Journal*, 46(2), 387-422.
- Honig, M. I., & Coburn, C. (2008). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, 22(4), 578–608.
- House bill 4627 of 2011, 96th Michigan State Legislature, Regular Session. (2011).
- Ingersoll, R. (1993). Loosely coupled organizations revisited. *Research in the Sociology of Organizations*, 11, 81–112.
- Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, 46(3), 587-613.
- Keesler, V. A., & Howe, C. (2016). Teacher Evaluation in Michigan. In J.A. Grissom & P. Youngs (Eds.), *Making the most of multiple measures: The impacts and challenges of implementing rigorous teacher evaluation systems* (pp.156-168). New York: Teachers College Press.
- Koedel, C., & Betts, J. R. (2007). Re-examining the role of teacher quality in the educational production function. Working Paper 2007-03. *National Center on Performance Incentives*.
- Kraft, M. A., & Gilmour, A. F. (2016). *Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness*. Retrieved from http://scholar.harvard.edu/files/mkraft/files/kraft_gilmour_2016_revisiting_the_widget_effect_wp.pdf?m=1456772152.
- Lavigne, A. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record*, 116(1), 1-29.

- Lowe Boyd, W., & Crowson, R. L. (2002). The quest for a new hierarchy in education: From loose coupling back to tight? *Journal of Educational Administration*, 40(6), 521-533.
- Lutz, F. W. (1982). Tightening up loose coupling in organizations of higher education. *Administrative Science Quarterly*, 27(4), 653-669.
- Major, M. L. (2013). How they decide a case study examining the decision-making process for keeping or cutting music in a K–12 public school district. *Journal of Research in Music Education*, 61(1), 5-25.
- Marsh, J. A., Kerr, K. A., Ikemoto, G. S., Darilek, H., Suttorp, M., Zimmer, R. W., & Barney, H. (2005). *The role of districts in fostering instructional improvement lessons from three urban districts partnered with the institute for learning*. Santa Monica, CA: RAND.
- Marzano, R. J., Toth, M., & Schooling, P. (2012). Examining the role of teacher evaluation in student achievement: Contemporary research base for the Marzano Causal Teacher Evaluation Model. Retrieved from http://www.marzanocenter.com/files/MC_White_Paper_20120424.pdf.
- McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, 9(2), 171-178.
- McLaughlin, M., & Talbert, J. (2003). *Reforming districts: How districts support school reform*. A Research Report. Document R-03-6. Seattle, WA: Center for the Study of Teaching and Policy.
- Measures of Effective Teaching Project. (2012). *Gathering feedback from teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.
- Measures of Effective Teaching Project. (2013). *Ensuring fair and reliable measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.
- Meyer, H. D. (2002). From “loose coupling” to “tight management”? Making sense of the changing landscape in management and organization theory. *Journal of Educational Administration*, 40(6), 515-520.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 340-363.
- Michigan Department of Education. (2014). *Educator Evaluations & Effectiveness in Michigan: An Analysis of 2013-2014 Educator Evaluation Systems Survey and Educator Effectiveness Data*, Lansing, MI: Author.
- Michigan Department of Education. (n.d). *Michigan Educator Evaluations Frequently Asked Questions (FAQs)*, Lansing, MI: Author.
- Milanowski, A.T. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, 79(4), 33-53.
- Murphy, J., & Hallinger, P. (1988). Characteristics of instructionally effective school districts. *The Journal of Educational Research*, 81(3), 175-181.

- Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation the case of the missing clothes? *Educational Researcher*, 42(6), 349-354.
- National Education Association (2015). Teacher evaluation. Retrieved from https://www.nea.org/assets/docs/20152_ESSA%20teacher%20evaluation.pdf
- Newmann, F. M., King, M. B., & Youngs, P. (2000). Professional development that addresses school capacity: Lessons from urban elementary schools. *American Journal of Education*, 108(4), 259-299.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237-257.
- Oosting (2015, October). Michigan teacher evaluation bill heads to Snyder after final approval in Senate. Mlive news. Retrieved from http://www.mlive.com/lansing/news/index.ssf/2015/10/michigan_teacher_evaluation_bi_1.html
- Orton, J. D., & Weick, K. E. (1990). Loosely coupled systems: A reconceptualization. *Academy of Management Review*, 15(2), 203-223.
- Park, V., & Datnow, A. (2009). Co-constructing distributed leadership: District and school connections in data-driven decision making. *School Leadership and Management*, 29(5), 477-494.
- Pogodzinski, B., Umpstead, R., & Witt, J. (2015). Teacher evaluation reform implementation and labor relations. *Journal of Education Policy*, 30(4), 540-561.
- Raywid, M. A., Schmerler, G., Phillips, S. E., & Smith, G. A. (2003). *Not so easy going: The policy environments of small urban schools and schools-within-schools*. Charleston, SC: ERIC Clearinghouse on Rural Education and Small Schools.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.
- 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. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review*, 261-266.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175-214.
- Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET Project*. Boulder, CO: National Education Policy Center. Retrieved from <http://nepc.colorado.edu/thinktank/review-MET-final-2013>
- Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11(1), 57-67.
- Shen, J., Gao, X., & Xia, J. (2016). School as a loosely coupled organization? An empirical

- examination using national SASS 2003-04 data. *Educational Management Administration & Leadership*. Advance online publication. <https://doi.org/10.1177/1741143216628533>
- Spain, A. K. (2016). Situating school district resource decision making in policy context. *American Journal of Education*, 122(2), 171-197.
- Spillane, J. P. (1996). School districts matter: Local educational authorities and state instructional policy. *Educational Policy*, 10(1), 63-87.
- Steinberg, M., & Donaldson, M. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, 11(3), 340-359
- Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy*, 27(4), 793-818.
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628–3651.
- Terry, K. (2010). We just can't seem to do what NCLB expects us to do: The case of an urban district focused on NCLB compliance. *Journal of Cases in Educational Leadership*, 13(1), 8-22.
- U.S. Department of Education. (2017) U.S. Secretary of Education Betsy DeVos announces release of updated ESSA consolidated state plan template. Retrieve from <https://www.ed.gov/news/press-releases/us-secretary-education-betsy-devos-announces-release-updated-essa-consolidated-state-plan-template>.
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative science Quarterly*, 21(1), 1-19.
- Weick, K. E. (1982). Administering education in loosely coupled schools. *Phi Delta Kappan*, 63(10), 673-676.
- Wohlstetter, P., Datnow, A., & Park, V. (2008). Creating a system for data-driven decision-making: Applying the principal-agent framework. *School Effectiveness and School Improvement*, 19(3), 239-259.
- Youngs, P., & Haslam, M. B. (2012). A review of research on emerging teacher evaluation systems. Washington, DC: Policy Studies Associates.

Essay 3: It is About the Culture: Early Career Teacher Turnover and Principal Leadership

Over the last decades, the U.S. teaching force has become less stable compared to other professions in terms of having higher levels of turnover (Ingersoll & Merrill, 2012). Several research studies have shown that such high turnover rates among teachers can have detrimental effects on school organizations (Allensworth, Ponisciak, & Mazzeo, 2009; Bryk & Schneider, 2002; Guin, 2004) and students' learning (Atteberry, Loeb, & Wyckoff, 2017; Ronfeldt, Loeb, & Wyckoff, 2013). Although addressing turnover issues among the entire teacher population is important, early career teacher (hereafter, ECT) turnover is particularly essential in a few respects. First, ECT turnover is closely related to general school staffing problems. As Ingersoll (2001) pointed out, school staffing problems result from a high level of teacher turnover, rather than supply-side deficits or increases in teacher retirement or student enrollment. More than 30% of new teachers leave the teaching profession within five years (Darling-Hammond & Sykes, 2003), and the turnover rate of early career teachers is significantly higher than that of experienced teachers (Allensworth et al., 2009; Kelly, 2004). That is, because high turnover rates are driven in part by ECTs, retaining more ECTs in the teaching profession and at their schools for longer periods may lower overall turnover rates, which is key to solving school staffing problems.

Second, retaining more ECTs can increase overall teaching quality. Given that years of experience in teaching is positively correlated with teachers' effectiveness, and improvement is significant in the first few years (Clotfelter, Ladd, & Vigdor, 2007; Rivkin, Hanushek, & Kain, 2005), retaining more ECTs and helping them to grow as teachers might contribute to enhancing overall teaching quality in school districts. Third, ECT turnover is linked to educational equity. ECTs are more likely to be hired to teach in hard-to-staff schools, which feature higher percentages of racial minority students and economically disadvantaged students, and are typically located in urban areas (Grissom, 2011; Hanushek, Kain, & Rivkin, 2004; Lankford, Loeb, & Wyckoff, 2002). That is, high levels of ECT turnover exacerbate issues in these already disadvantaged schools, and this contributes to achievement gaps among students in different districts.

Researchers have focused on various factors that affect teachers' retention decisions and actual turnover. In most cases, when teachers leave, they choose another position over their current position, despite their previous investment to achieve the position. The question here is which factors might affect the relative attractiveness of the current position? There is some agreement among researchers about determinants of teacher turnover in terms of teachers' individual characteristics (e.g., Borman & Dowling, 2008; Clotfelter, Ladd, Vigdor, & Diaz, 2011), student composition (e.g., Boyd, Lankford, Loeb, & Wyckoff, 2013; Hanushek et al., 2004), school resources and structure (e.g., Allensworth et al., 2009), teachers' salary (e.g., Ingersoll & May, 2012) and the organizational context of schools (e.g., Boyd, Grossman, Ing, Lankford, Loeb, & Wyckoff, 2011; Ladd, 2011). However, it is clear that teachers' individual characteristics and/or student composition are difficult to change, and enhancing school resources and/or teacher salaries is not easy to accomplish in a short time. At the same time, many recent research studies have shown that some aspects of school organization amendable by policy can impact teacher turnover even after controlling for less malleable factors (e.g., Ingersoll & May, 2011; Ladd, 2011; Loeb, Darling-Hammond, & Luczak, 2005). In particular, several research studies have shown that principal leadership affects teachers' planned retention decisions as well as their actual turnover rates (Allensworth et al., 2009; Boyd et al., 2011; Grissom, 2011; Ingersoll & May, 2012; Ladd, 2011; Pogodzinski, Youngs, Frank, & Belman, 2012; Youngs, Kwak, & Pogodzinski, 2015). Although these studies provide important insights about principals' roles in shaping teachers' working conditions and affecting teacher turnover, there are some gaps in the literature.

First, although principal leadership is multi-faceted, most quantitative studies have used a broad definition of leadership to understand the factors that affect teacher turnover. As Grissom, Loeb, and Master (2013) argued, however, aggregating principals' behavior may make it difficult to attain a more nuanced understanding of principals' influence on teachers. Second, as Simon and Johnson (2013) noted, previous studies based on quantitative analyses mostly draw on one or two years of data (e.g., Boyd et al., 2011; Grissom, 2011; Ingersoll & May, 2011; Johnson, Kraft, & Papay, 2012; Kelly & Northrop, 2015), while it is important to examine this issue with a longer-term perspective. As teacher development

theories have shown, teacher needs vary across different developmental stages (Berliner, 1988; Fuller, 1969; Kagan, 1992) and, thus, the same factors may have different effects on teacher turnover as teachers gain experience. Some factors may affect teacher turnover for one or two years, while other factors may have longer effects on teacher turnover. Moreover, analyses of factors that affect teacher turnover would be more robust and reliable if longitudinal data were to be used.

This study aims to better understand the impact of principal leadership on ECT turnover by addressing these gaps. I draw on five years of data from a nationally representative dataset collected by the National Center for Education Statistics (NCES): the Beginning Teacher Longitudinal Study (BTLS). I apply a discrete time survival analysis to integrate longitudinal data on principal leadership with teacher turnover information. In particular, following the framework of Youngs and colleagues (2015), I examine the impact on ECT turnover of three different aspects of principals' leadership (i.e., instructional leadership, leadership related to managing student behavior, and leadership related to fostering supportive school culture), as well as general leadership behaviors.

The term "teacher turnover level" in this essay refers to two different outcomes. The first part of the analysis focuses on any form of leaving teachers' original schools; that is, stayers vs. movers (i.e., those who move to a new school) or leavers (i.e., those who leave the profession). The second part of the analysis focuses on leaving the profession altogether (i.e., stayers and movers vs. leavers). At the school level, the first outcome is important because when a teacher leaves a school, it does not make a difference to the school whether the teacher moves to another school or leaves the profession altogether, as the school loses a member of its staff and must fill the vacancy (Ingersoll, 2001). In contrast, at the level of the entire teaching force, the latter is important because when a teacher does not leave the profession but moves to another school, the teacher will still grow as a professional, and it will not be necessary to train a new teacher to replace him/her in the field. In other words, examining the first outcome is more closely related to immediate results at the school organizational level, while examining the second outcome is related to the longer-term influence on the teaching force as a whole.

The possible impact of different aspects of leadership (i.e., instructional leadership, leadership related to managing student behavior, and leadership related to fostering school culture) measured at Year 1 and general principal behaviors measured at Years 3 to 5 on these outcomes are analyzed in separate models. After analyzing the main effects of principal leadership on ECT turnover, I examined effects of interactions between principals' leadership and different aspects of school context (i.e., school location, charter vs. public schools, and elementary vs. secondary schools) and time indicators. For example, it is possible that principal leadership has a different level of impact as ECTs gain experience, or it has a stronger effect on ECT turnover in schools with high percentages of low-SES (socio-economic status) students, as Ladd (2011) and Grissom (2011) showed.

The research questions for this essay include:

1. How are ECTs' perceptions about principal leadership, in terms of three aspects of leadership and general principal behavior, associated with their turnover levels (i.e., leaving the school and leaving the profession)?
2. How does the association between ECTs' perceptions about principal leadership and their turnover levels vary across years?
3. How does the association between ECTs' perceptions about principal leadership and their turnover levels vary based on different aspects of school context?

From a practical perspective, this study can inform ways to support ECTs and their principals in enhancing teacher retention, while from a theoretical standpoint, this paper uses longitudinal data to explore different aspects of principal leadership and their potential outcomes. The next section of this essay reviews the relevant literature; the third section presents the theoretical framework of the study; and the fourth section introduces the data and analytical approach. Finally, I present results about the impact of principal leadership on ECT turnover based on a discrete time survival analysis, followed by a discussion focused on policy implication, limitations of the study, and implications for future research.

Factors That Affect Teacher Turnover

Some studies have focused specifically on ECT turnover and various factors that affect such decisions (e.g., Johnson & Birkeland, 2003; Kelly & Northrop, 2015; Smith & Ingersoll, 2004; Youngs et al., 2015). In addition to such literature, I also reviewed research studies that examined turnover among teachers with various experience levels in order to develop a comprehensive picture of the factors that affect turnover.

In order to become a teacher, one has to invest a considerable amount of resources (i.e., time and money for training) as well as opportunity costs that he/she would have earned from working in other professions. This also applies to teachers' adjusting themselves to a school; when a teacher starts to work at a new school, he/she needs to spend time and energy to become familiar with the school environment. That is, if a teacher decides to leave the profession or the current school, it would mean that their difficulties or dissatisfaction was so great that he/she decided to sacrifice the resources that they have spent to achieve their current position.

Earlier studies about teacher turnover have focused on who might face a high level of difficulty in teaching and/or who might be sensitive to such difficulty, while more recent studies have focused on factors contributing to such difficulty (Simon & Johnson, 2013). According to studies of the association between teacher characteristics and turnover, ECTs are more likely to leave their schools than experienced teachers (Allensworth et al., 2009), and teachers who are male, white, or married are more likely to leave their current schools than teachers who are female, non-white, or single, respectively (Borman & Dowling, 2008). In other words, working in a given teaching position might require early career, male, white, and married teachers to endure greater challenges than it does for their counterparts, or they are more sensitive to such challenges. For example, ECTs who usually lack instructional expertise might face a higher level of emotional challenges in their teaching compared to experienced teachers. Male teachers and married teachers are likely to be more sensitive to low salaries. While researchers generally agree on the impact of teachers' demographic backgrounds on turnover, studies of teacher quality and turnover have produced contrasting results depending on their measures of teacher quality.

Boyd and colleagues (2008), and Henry, Bastian, and Fortner (2011) showed that among first-year teachers in New York City, teachers who were less effective in terms of improving students' achievement scores were more likely to leave their schools. In District of Columbia Public Schools (DCPS), the quality of entering teachers was systematically higher than that of exiting teachers based on the district teacher evaluation system, which includes value-added measures as well as observations (Adnot, Dee, Katz, & Wyckoff, 2017). In addition, based on a meta-analysis of factors that affect teacher turnover, Borman and Dowling (2008) found that non-certified teachers had a high possibility of leaving their current schools. However, these findings about teachers' effectiveness are not consistent with studies showing that teachers with stronger academic backgrounds are more likely to leave their schools (Borman & Dowling, 2008; Boyd, Lankford, Loeb, & Wyckoff, 2005; DeAngelis & Presley, 2007; Kelly & Northrop, 2015; Redding & Smith, 2016), and National Board Certified teachers are more likely to leave their school districts (Goldhaber & Hansen, 2009). These inconsistent results can be explained by the fact that low-performing teachers might face more challenges in their teaching, but high-performing teachers also might be more sensitive to the challenges of their work since they may have more alternative job options based on their certification or strong academic backgrounds.

In addition to teacher characteristics, student composition also plays a significant role in teacher turnover in that it can affect the challenges that teachers encounter. A substantial body of literature documents the particularly high levels of teacher turnover in schools that serve high percentages of racial/ethnic minority, low-SES students, and/or low-performing students (Allensworth et al., 2009; Boyd et al., 2005; Boyd et al., 2013; Clotfelter et al., 2011; Hanushek et al., 2004; Ingersoll, 2001; Ingersoll & May, 2012; Scafidi, Sjoquist, & Stinebrickner, 2007). For example, using data from the 2003-04 Schools and Staffing Survey and the 2004-05 Teacher Follow-Up Survey, Ingersoll and May (2012) examined mathematics and science teachers' turnover and showed that, in general, teachers moved from schools that serve low-income, minority, and/or urban students to schools that serve higher-income, non-minority, and/or suburban students. The authors also noted that leaving the profession follows the same pattern. Allensworth and colleagues (2009) used multiple years of teacher turnover data in Chicago Public

Schools and reported that 100 schools that suffered from chronic teacher mobility were serving predominantly low-income African-American students with low achievement scores. In another study, student composition was seen as more influential than other factors, such as salary (Hanushek et al., 2004). A more serious problem is that teachers with strong pre-service qualifications, more experience, and higher scores on certification exams were more responsive to student composition or student achievement compared to other teachers, in terms of retention decisions (Boyd et al., 2005; Clotfelter et al., Ladd & Vigdor, 2011; Lankford et al., 2002).

Recent studies about teacher turnover, however, suggest another story underlying such patterns: “(T)eachers who leave high-poverty schools are not fleeing their students. Rather, they are fleeing the poor working conditions that make it difficult for them to teach and for their students to learn” (Simon & Johnson, 2013, p. 1). In fact, school characteristics, including student composition, often become less important for teacher turnover when other school context variables, such as principal leadership and different aspects of teachers’ working conditions, are included in research studies (DeAngelis & Presley, 2007; Johnson et al., 2012; Loeb et al., 2005; Simon & Johnson, 2013). That is, teachers tend to leave those schools because of poor working conditions rather than because they are asked to teach certain students.

Among school organizational factors, principal leadership has been studied extensively by researchers as one of the most significant aspects of school context for teacher turnover. Ladd (2011) analyzed statewide teacher survey data in North Carolina to examine the relationship between teachers’ perceptions about their working conditions and their turnover. The study used a broad measure of teachers’ perceptions about their principal leadership, including leadership related to maintaining discipline in the classroom, trusting teachers, and teachers’ involvement in decision-making. Among various aspects of school organizations, the quality of principal leadership showed the strongest impact on teachers’ intended and actual turnover. The impact of principals was independent from student composition in each school, and it was more salient in schools that served many minority students. Grissom (2011) found a similar result using nationally representative data, the 2003-04 School and Staffing Survey (SASS) and

2004-05 Teacher Follow-up Survey (TFS). He reported that the positive impact of principal effectiveness on teacher retention was stronger in disadvantaged schools. Grissom (2011) also used a single aggregated measure of principal leadership based on six survey items, such as “The principal knows what kind of school he/she wants and has communicated it to the staff” and “The school administration’s behavior toward the staff is supportive and encouraging.”

Combining survey data from all first-year teachers in New York City, Boyd and colleagues (2011) also showed that the support of administrators had a significant impact on teachers’ intended and actual turnover. In their analysis, dissatisfaction with the administration was a more important factor for leaving or planning to leave than dissatisfaction with students’ behavior. The administration variable in this study was measured in a similar way with the previous studies, but principal behavior towards external pressure and teacher evaluation were also included. Johnson and colleagues (2012) found similar results using teacher survey data from Massachusetts. Principal leadership measures were included in the analysis along with nine other measures of teacher working conditions. The authors measured principal leadership based on teachers’ responses about principals’ feedback on instruction, their ability to create an orderly and safe instructional environment, and whether they addressed teachers’ concerns about issues in the school. The authors included school culture as a separate variable, measured by teachers’ perceptions about “the extent to which the school environment is characterized by mutual trust, respect, openness, and commitment to student achievement” (p. 14). The authors showed that these two factors along with other teacher working conditions are much more important than other physical resources, such as more time and the nature of school facilities.

Taken together, it is clear that principal leadership has a significant influence on teacher turnover. This is not surprising in that strong principal leadership can reduce challenges that teachers at the school might experience in fulfilling their responsibilities. However, as noted earlier, various facets of principal leadership and their potentially different effects on teacher turnover have not been fully examined in most studies. As Simon and Johnson (2013) argued, “these studies do not closely analyze what it is about school leadership that matters...” (p. 15). In order to provide specific policy recommendations, examining

which aspects of principal leadership have a strong association with teacher turnover is essential. Accordingly, this study distinguishes among three aspects of principal leadership and examines how they might shape teacher turnover differently. In addition, it has been shown that school size, teachers' involvement in decision making, teacher salary, and staff relations have significant associations with teacher turnover (Allensworth et al., 2009; Ingersoll, 2001; Ingersoll & May, 2011; Simon & Johnson, 2013). In order to understand the impact of principal leadership on ECT turnover, I included those factors as control variables for the analysis. Based on this review of relevant literature, the next section presents the theoretical framework underlying this study.

Theoretical Framework

The goal of this study is to understand how ECT turnover levels are associated with their perceptions about principal leadership. In order to understand the mechanisms of how principals can influence teacher turnover levels, I modified the framework of Youngs and colleagues (2015). I conceptualize principal leadership by focusing on three different, but possibly related, aspects: instructional leadership, leadership related to managing student behavior, and leadership related to creating a supportive school culture.

Instructional leadership studies originated in the 1980s' effective schools movement (Hallinger, 2003; Marks & Printy, 2003; Murphy, 1988). Principals working in "effective schools" were described as "culture builders," who set high academic expectations for all students and teachers, and as "goal-oriented," who established clear goals for their schools and encouraged teachers to work to achieve the goals (Hallinger, 2005, pp. 223-224). Such principals are more likely to provide ECTs with sufficient learning opportunities and useful feedback on their instruction, while "weak instructional leaders typically fail to bridge beginning teachers' current curricular and pedagogical knowledge with broader perspectives" (Youngs et al., 2015, p. 167). By working with strong instructional leaders, ECTs are more likely to develop expertise and experience fewer challenges; this can enhance their retention rates.

A second important aspect of principal leadership is leadership related to managing student behavior. Student behavioral issues not only have a direct impact on teacher turnover as an important part

of teachers' working conditions (Allensworth et al., 2009; Boyd et al., 2011; Johnson et al., 2005; Ladd, 2011), but also an indirect impact on teacher turnover by impeding teachers' instructional practice. In fact, ECTs typically have difficulties in managing students, and they lack systematic support for this aspect of their work (Johnson & Birkeland, 2003). Given that ECTs have limited expertise, principals' active support for managing student behavioral issues can have a significant impact on teacher retention levels. That is, principals who demonstrate strong leadership related to managing student behavior might help ECTs to focus on their instructional practice with less distraction, and ECTs' general working conditions can be improved by having fewer challenges in their classrooms (Youngs et al., 2015). Moreover, by modeling how to deal with such problems, principals can provide professional learning opportunities for teachers.

A third key aspect of principal leadership is their leadership related to creating a supportive professional culture in schools. It is well known that school administrators have an enormous influence on school cultures among teachers (Pogodzinski et al., 2012; Rorrer & Skrla, 2005; Supovitz, Sirinides, & May, 2009). As formal leaders of school organizations, school administrators can promote or discourage certain school cultures, such as trust among school community members (Bryk & Schneider, 2002) and a professional learning community (Louis, Marks, & Kruse, 1996). School-based opportunities to learn from other teachers, based on mentoring programs and collaboration time among teachers, have been found to have a significant positive effect on teacher retention (Allensworth et al., 2009; Borman & Dowling, 2008; Smith & Ingersoll, 2004). Principals can promote school-based learning through various strategies ranging from allocating resources for common lesson planning or mentoring to establishing a supportive atmosphere among teachers. In this case, ECTs can increase their expertise and share their concerns with other teacher colleagues, which may enhance their retention levels. Thus, it is important to take into account not only principals' direct support for each teacher, but also their indirect support for the whole school by fostering a supportive culture. It should be noted that these three aspects of principal leadership are defined broadly in this paper; they include not only principals' direct behavior but also aspects of teachers' work that might be affected by principals' behavior. For example, whether ECTs

received certain support, such as a reduced teaching schedule, was included in the instructional leadership component because principals can intentionally allocate additional resources to provide support for ECTs.

The final part of this framework are the three different turnover results for ECTs: staying at the same school (i.e., stayers), moving to other schools (i.e., movers), and leaving the profession altogether (i.e., leavers). Based on a disruptive perspective on the effects of teacher turnover, movers and leavers have an equally detrimental effect on school organizations, since both result in discontinuity issues in a given school organization and cost resources to recruit new teachers (Ronfeldt et al., 2013). However, given the large amount of resources needed to train new teachers and students' lost opportunity to learn from more experienced teachers, ECTs leaving the profession can cause severe problems in the long run. Teachers' motivation for leaving the school and leaving the profession can also vary; teachers may move to other schools because they are not satisfied with their school organization, district, or even state system, but are satisfied with the daily job itself. Teachers may leave the profession because they are not satisfied with the teaching job itself or do not have the proper skills to fulfill their responsibilities. In the current study, the focus is on principal leadership; it is possible that such leadership is only related to teachers leaving the school, but not to teachers leaving the profession, or vice versa. If principal leadership only has a significant association with teachers leaving the school, then it is important to explore other ways to address the issue of teachers leaving the profession. In this sense, this study can provide useful policy implications by differentiating teacher turnover into different categories. Based on this framework, I introduce the data and analytic approach in the following section.

Method

Data

In order to understand the association between principal leadership and ECT turnover levels, I drew on a nationally representative dataset collected by the National Center for Education Statistics (NCES) — the Beginning Teacher Longitudinal Study (BTLS). The BTLS was initially part of the Schools and Staffing Survey (SASS) in 2007-08 and followed the same set of first-year teachers for five years. The structure and content of the survey data is similar to the SASS data, which has been described

as “the largest and the most comprehensive data source” on U.S. teachers (Ingersoll & May, 2012, p. 440). There were 1,990 eligible teachers included in the final BTLS sample, and the response rate ranged from 91.4% to 77.7% through the five waves (Burns, Wang, & Henning, 2011). All data are self-reported by teachers and/or principals, and it was difficult to use any aggregated measures at the school level because 69% of the participating teachers were the only ECTs in their schools. I drew on all five years of BTLS data in order to obtain data about principal leadership, teacher characteristics, and teacher employment status, and I drew on the 2007-08 SASS principal survey data for school characteristics. Figure 1 summarizes the timeline of BTLS and SASS data collection. Principal data from SASS was not used based on their own perceptions about their leadership not having any significant association with ECT turnover. In addition, since most ECTs worked with several different principals during their first five years, limiting the sample to ECTs whose principals stayed at the same school decreased the sample size significantly. Thus, I did not consider whether ECTs worked with the same principals for multiple years for this study.

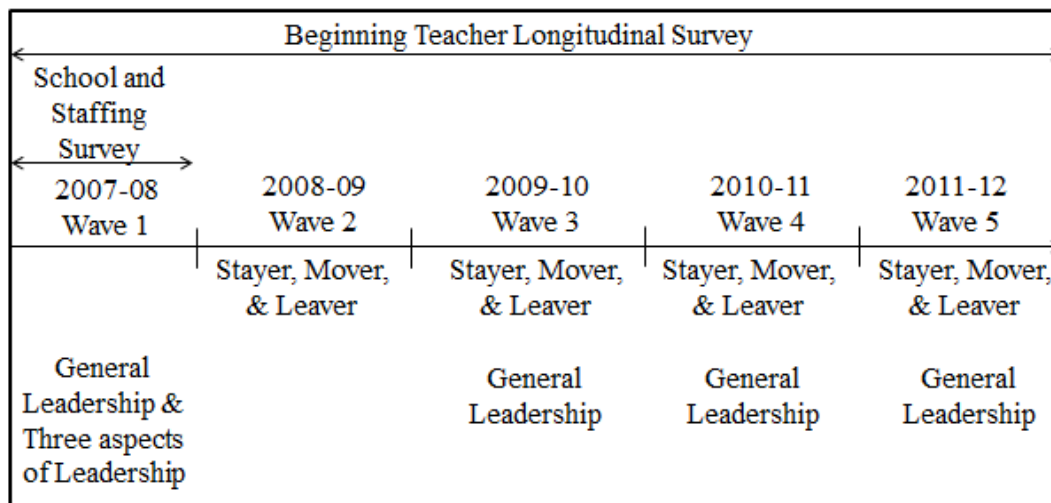


Figure 6. Time Line of BTLS Data Collection

Measures

ECTs leaving the school and leaving the profession. The main dependent variable is ECTs’ employment status reported during the following school year. For example, the 2008-09 BTLS survey

tracked sampled ECTs who were first-year teachers in 2007-08 and asked them to report whether they 1) stayed at the same school, 2) moved to another school, or 3) left the teaching profession altogether. Based on their response, they were categorized as 1) stayers, 2) movers, or 3) leavers. For my study, these three groups are regrouped into two categories; whether the teacher left their school and whether they left the profession. For leaving the school, the teacher was assigned a 1 if they left their school (i.e., movers or leavers), and a 0 if they stayed at the same school (i.e., stayers). For leaving the profession, the teacher was assigned a 1 if they left the profession altogether (i.e., leavers), and a 0 if they stayed at the school or moved to another school (i.e., stayers or movers). These two variables, leaving the school and leaving the profession, were included as dependent variables for separate models.

ECTs' perceptions about principal leadership. The principal leadership variables were divided into two categories: time-invariant measures of three specific aspects of leadership (i.e., instructional leadership, leadership related to managing student behavior, and leadership related to creating a supportive school culture) and a time-variant measure about principals' general behavior. As part of the 2007-08 SASS survey, the BTLS survey provided rich data about ECTs' perceptions about their principals. I included three aspects of principal leadership measured by the 2007-08 survey as time-invariant variables for some models in order to estimate the effects of different aspects of principal leadership on teacher turnover. In terms of instructional leadership, I took the standard mean of six items about whether teachers had received various supports for their professional learning from their schools. For instance, teachers answered items about whether they had support for "common planning time with teachers in your subject," and "extra classroom assistance" (NCES, 2008, p. 25, $\alpha=0.46$). Principals' leadership related to managing student behavior was calculated based on teachers' responses to six items, such as "the level of student misbehavior in this school interferes with my teaching" and "my principal enforces school rules for student conduct and backs me up when I need it" ($\alpha=0.689$). Principal leadership related to a supportive culture was calculated based on teachers' responses to six items such as "the school administration's behavior toward the staff is supportive and encouraging" and "the principal

knows what kind of school he or she wants and has communicated it to the staff” ($\alpha=0.802$). See Appendix A for the list of all survey items used for the analysis.

In terms of time-variant principal leadership, I focused on ECTs’ perceptions about principals’ general behavior. While the third, fourth, and fifth waves of the BTLS survey did not include detailed questions about multiple aspects of principal leadership as the first wave did, those surveys asked ECTs about their general judgment about principals’ behaviors. To maintain consistency among measures, I limited the principal leadership items from the Year 1 survey to three items asking directly about principal’s behaviors rather than general support available for ECTs; whether they have “Regular supportive communication with your principal, other administrators, or department chair,” their response to “My principal enforces school rules for student conduct and backs me up when I need it,” and “The school administration’s behavior toward the staff is supportive and encouraging” ($\alpha=0.704$). For the third, fourth, and fifth waves of the BTLS survey, I calculated this variable based on the mean of the responses to the seven items, such as “My principal supports me in classroom management issues when I need it,” “My principal supports me in my interactions with parents when I need it,” and “My principal is approachable” ($\alpha=0.928$ for 3rd year, $\alpha=0.926$ for 4th year, and $\alpha=0.925$ for 5th year). Since the 2nd-year survey did not have any items about principal leadership, I used the Year 1 principal variable for Year 2 as well.

Control variables. Control variables fall into two categories: individual teacher background variables and other school context variables, besides principal leadership. These variables were collected based on teachers’ data in the BTLS survey and principals’ data in the SASS survey from 2007-08. First, various teacher background variables were included. Informed by existing literature (e.g., Hanushek et al., 2004; Loeb et al., 2005), I included the following as teacher-level control variables: gender, race, whether the teacher had an alternative teaching certificate, salary, working hours per week, teaching subject, teacher preparation (i.e., both degrees and teachers’ perceptions about their preparation), autonomy, whether the teacher experienced formal induction, union membership, whether the teacher is highly qualified (i.e., HQT), the level of commitment, and hours they allocated to professional development. All

teacher characteristics, except salary (logged) and HQT, were measured based on the Year 1 BTLS survey. Following Redding and Smith (2016), I included subject taught as a dummy variable equal to 1 if a teacher taught an in-demand subject; i.e., mathematics, science, special education, or English as a second language.

Teachers' perceptions about their preparation were calculated based on their responses to six items under the question, "In your first year of teaching, how well prepared were you to...", such as "handle a range of classroom management or discipline situations," "use a variety of instructional methods," and "teach your subject matter" ($\alpha=0.814$). Teacher autonomy was measured by six items under the question, "How much actual control do you have in your classroom at this school over the following areas of your planning and teaching?", such as "selecting textbooks and other instructional materials," "selecting content, topics, and skills to be taught," and "selecting teaching techniques" ($\alpha=0.706$). Teacher commitment is one of the most important control variables in order to estimate precisely the effects of the principal leadership on teacher turnover and leaving the profession, in that it can control for teachers' attributes that might affect their own career trajectory. The variable was calculated based on teachers' responses to three items: "If I could get a higher paying job I'd leave teaching as soon as possible," "I think about transferring to another school," and "How long do you plan to remain teaching?" ($\alpha=0.526$). This commitment variable was measured prior to the general principal leadership variable, except for the Year 1 data, while it was measured at the same time with the three aspects of principal leadership. Professional development was calculated based on the mean of the hours that teachers had spent for the past 12 months on professional development training about specific content, using computers for instruction, reading instruction, student discipline and classroom management, teaching students with special needs, and/or teaching students with limited-English proficiency.

Second, school context variables were included: number of students, percentage of racially minority students, whether the school made adequate yearly progress (AYP), school safety, parents' involvement, percentage of free- and reduced-price-lunch eligible students, school location, whether the school was a charter school, and school level (e.g., whether the school was an elementary school or high

school). Among these variables, AYP status, school safety, and parents' involvement were obtained from the 2007-08 SASS principal survey responses. Only percentage of free- and reduced-lunch eligible students, school location, and whether the school was a charter school were available as time-variant variables in the BTLS data. Other than these variables, all school-level control variables were obtained from the 2007-08 BTLS data. The school safety variable was calculated based on 13 items from the principal survey, under the question, "To the best of your knowledge, how often do the following types of problems occur at this school?" The items included "physical conflicts among students," "robbery or theft," "vandalism," "student use of alcohol," and "student use of illegal drugs" ($\alpha=0.847$). Parents' involvement was calculated based on principals' responses about the question, "Last school year (2006-07), what percentage of students had at least one parent or guardian participating in the following events?" The events included "open house or back to school night," "all regularly scheduled schoolwide parent-teacher conferences," "one or more special subject area-events (e.g., science fair, concerts)," and "volunteer in the school on a regular basis" ($\alpha=0.818$). All variables were calculated by taking a standard mean of related items and Appendix A contains a detailed list of the survey items that I used.

Analytical Approach

In order to understand the association between principal leadership and ECT turnover levels across years, I applied a discrete-time survival analysis framework. Survival analysis, originating from medical research, has been widely applied in different disciplines; the concept of survival has expanded to a broader scope (Liu, 2012) and in educational research includes students' dropout rates (Lesik, 2007; Murtaugh, Burns, & Schuster, 1999; Plank, Deluca, & Estacion, 2008), failure on tests (Schultz, Evans, & Serpell, 2009), and teacher turnover levels (Redding & Smith, 2016). This analysis is advantageous compared to logistic regression in that it examines the event occurrence itself across multiple years and takes into account the censored nature of the data (Singer & Willett, 1993). As noted earlier, there are two types of "survival" in this study; 1) teachers stay at the same school (i.e., they never move to another school nor leave the whole profession) and 2) teachers stay in the teaching profession (i.e., they stay at the same school or move to a different school).

I started by conducting a univariate analysis of ECTs' survivor function by principal leadership quartiles. I divided ECTs into four groups based on their perceptions about principal leadership in terms of instructional leadership, leadership related to student management, and leadership related to creating a supportive culture measured at Year 1 (i.e., in 2007-08). Since univariate analysis aims to show descriptively whether different groups had different survival functions, I used quartiles instead of using the raw values of ECTs' perceptions of principal leadership, in order to have a fewer number of groups. Kaplan-Meier survival curves and the Cox regression-based test for equality of survival curves were used (Hosmer, Lemeshow, & May, 2008). Since the general leadership variable is time-variant, it was hard to apply these univariate approaches. Next, I proceeded to the discrete-time survival analysis (Singer & Willett, 1993). Although the Cox proportional hazard regression model has been commonly used to analyze survival data (Liu, 2012), there are a few advantages of using discrete-time survival analysis in this context. First, the data were collected only once each year, so the continuous time frame of the Cox model is not applicable; second, the general principal leadership variables in this study are time-variant predictors for some models, while the Cox model assumes that predictors are consistent over time (Singer & Willett, 1993).

In order to answer the first research question, about the association between principal leadership and teacher turnover, the first model is set as follows:

$$\Pr(\text{Leaveschool}_{ij}) = \alpha_1 T_{2ij} + \alpha_2 T_{3ij} + \dots + \alpha_4 T_{5ij} + \beta_1 (\text{Principal leadership}_i)^1 + \mathbf{X}_i \gamma + \mathbf{A}_{ij} \delta + e_{ij} \quad (1)$$

Where, $\Pr(\text{Leaveschool}_{ij})$ is the probabilities that teacher i left the school (i.e., moved to another school or left the profession altogether) in year j . These outcome variables are equal to 0 until the event occurs and become 1 when the event occurs. After the year the event occurred, the variables appear as a missing value (i.e., censored). That is, if a teacher left the original school at Year 3, the Year 2 outcome is 0, and the Year 3 outcome is 1, and it would be missing for Years 4 and 5. In this study, I did not consider returners (i.e., teachers who left the school/profession and came back later years), since only about 100 teachers were categorized as returners across five years, and it is hard to apply a survival analysis for those returners. T_2 through T_5 are indicators for each year; I started the analysis at T_2 because this was

the first year for which turnover information was available. The Principal leadership variable includes three aspects of principals' leadership measured at the Year 1, and they were included separately, due to a multicollinearity concern. $\mathbf{X}_i$ includes time-invariant and $\mathbf{A}_{ij}$ includes time-variant control variables: school and teacher characteristics. The same model was applied to $\text{Pr}(\text{LeaveProf}_{ij})$, which is the possibilities that teacher i left the profession (i.e., left the profession altogether) in year j . After analyzing the three aspects separately, the coefficients of each aspect were compared to each other based on Hausman test (Cohen & Cohen, 1983).

For the second model, I included principal leadership as a time-variant variable. The second model is as follows:

$$\text{Pr}(\text{Leaveschool}_{ij}) = \alpha_1 T2_{ij} + \alpha_2 T3_{ij} + \dots + \alpha_4 T5_{ij} + \beta_1 (\text{Principal leadership}_{ij-1}) + \mathbf{X}_i \gamma + \mathbf{A}_{ij} \delta + e_{ij} \quad (2)$$

That is, a teacher leaving the school at a given time point is a function of their principal leadership during the previous year. All other control variables remained the same. The same model was applied to $\text{Pr}(\text{LeaveProf}_{ij})$, which is the probabilities that teacher i left the profession (i.e., left the profession altogether) in year j as well. In various models, I included a two-year lagged principal leadership variable along with a one-year lagged effect, in order to examine whether principal leadership has a lagged effect on teacher turnover for more than a year later.

In order to answer the second research question, I included terms for the interactions between time and principal leadership in equations (1) and (2). It is possible that principal leadership has a different level of influence on ECT turnover as ECTs gain experience. For the third research question, terms for the interactions between principal leadership and school level (i.e., elementary or not and high school or not), principal leadership and location of the school (i.e., suburban or not), and principal leadership and whether the school was a charter school were included. This analysis examines potential heterogeneous effects of principal leadership on teacher turnover across years in different school contexts. In order to avoid multicollinearity, the variables used for calculating the interaction terms were grand-mean centered.

It is critical to note the issue of bias in estimating the association between principal leadership and ECT turnover. As Boyd and colleagues (2011) pointed out, teachers' self-reported intention of leaving their schools can be inaccurate. More importantly, teachers' perceptions about their working conditions and principal leadership can be imprecise measures of actual working conditions and leadership in that teachers who consider leaving the school are more likely to report negative perceptions about their schools. In this case, such unobservable attributes of teachers might distort the association between principal leadership and teacher turnover. In order to address these issues, I use three approaches. First, I only focus on actual turnover, rather than teachers' intention in the analysis, since teachers' reports can be inaccurate. Second, I include teachers' intention to leave the school or the profession (i.e., commitment) measured at Year 1 as a control variable in models. That is, I examine the impact of the principal leadership on the turnover of teachers whose commitment level was the same in their first year. Third, I quantified the robustness of the inference by using Robustness Indices (Frank, 2000; Frank et al., 2008; Frank et al., 2013). This shows how much of the sample would need to be replaced in order to invalidate the inference.

In terms of weights, the BTLS data provide three sorts of sampling weights and replicate weights for each year (i.e., analysis replicate weights, retrospective analysis weights, and longitudinal analysis weights). The sampling weights were calculated based on the inverse of probability of being sampled and nonresponses at the school level. Since BTLS was not based on a simple random sample, but a stratified probability proportionate-to-size sample, it is necessary to use sampling weights to achieve accurate results (Kaiser, 2011). However, there are two issues in using the sampling weights in the data: 1) using weights from different years and 2) highly skewed weights. First, given the survival analysis framework, it is not clear which years should be used in selecting weights. Using a longitudinal weight in Year 5 for all teachers may not solve this issue because some teachers had already disappeared in the data before Year 5 if they left their school or the profession before then but completed earlier surveys.

For example, if a teacher left their school and moved to another school in Year 3, the teacher would have missing values in Year 4 and Year 5 when I conducted the analysis on ECTs leaving the

school. Thus, applying Year 5 longitudinal weights is not appropriate for this teacher. In order to address this issue, I created new sampling weights for leaving the school and leaving the profession separately based on the time point when each teacher appeared in the data for the last time. For example, when a teacher left the school and moved to another school in Year 3 and stayed at a new school for Years 4 and 5, the leaving the school weight for this teacher is a Year 3 longitudinal weight. However, this teacher did not leave the profession until Year 5, so the leaving the profession weight is a Year 5 longitudinal weight. In doing so, I could apply the appropriate weights for each teacher.

Another issue was related to the wide range and severe skewness of the weights. Based on the new leaving the school weights, the maximum is 2,031.577 and the mean is 85.545. In order to identify any systematic patterns in teachers' and schools' characteristics and weights, I report the correlations among them [See Appendix B]. The overall correlation was negligible (i.e., $r < 0.15$), but elementary school teachers tended to have a higher level of weights (i.e., $r = 0.261$ for the leaving the school weight, and $r = 0.268$ for the leaving the profession weight), and high school teachers tended to have a lower level of weights ($r = -0.1946$ for the leaving the school weight, and $r = -1.999$ for the leaving the profession weight). The patterns in the correlations are consistent across the two kinds of weights.

The second issue was about the skewness of the weights. While only 10 out of 1,990 teachers had more than 1,000 as their weights, a number of teachers had less than 100. This is problematic in that this small number of teachers can influence all of the results. Therefore, I transformed these to the natural log of the weights. Since 0 was a meaningful value in the original weights due to the missing values, I kept 0s by adding 1 and transforming the weights into the natural log. After the transformation, the distribution was close to a normal distribution. As a robustness check, I also conducted the same survival analysis using the original weights and excluding any weights to check whether the results are sensitive to different weights. In addition, without setting a sampling weight, I ran the survival model with terms for the interaction between the weights and principal leadership variables using individual teacher-level clustered errors to see whether the impact of principal leadership varied by the weights. As a result, the

coefficients of the interaction terms between the weights and the principal leadership variables were all close to zero.

Based on preliminary analysis regarding sample weights, school size, percentage of minority students, and teachers' race seemed to be key variables for the sampling weights. Accordingly, I included terms for the interactions between those three key variables and principal leadership without weights to see if including those interaction terms would capture the effects of weighting. Including those interaction terms did not make much difference; the results were close to the ones from the models without any weights, so they generally did not capture the effects of using weights. The results related to the weights are reported in Appendices C and D.

Another feature of the BTLS data is the option of using replicate weights. Since "direct estimates of sampling errors that assume a simple random sample will typically underestimate the variability in the estimates," it seemed that using replicate weights would make sense (Kaiser, 2011, p. B-14). The issue for this study is that the survival analysis involves multiple measures from each teacher, so it is important to use cluster-adjusted standard error at the teacher level. However, using replicate weights based on the balanced repeated replication technique, suggested by NCES, is not compatible with the cluster-adjusted standard errors in STATA. Given the survival analysis framework, the teachers in the data are not consistent across years, so it is hard to apply a balanced repeated replication. Therefore, for the main analysis, I used standard errors clustered at the teacher level. The results from the analysis using replicate weights (i.e., Year 5 longitudinal replicate weights) with a balanced repeated replication technique are reported in Appendix E.

Results

Descriptive univariate analysis

Table 20 presents descriptive statistics of the variables used for the analysis. Both weighted and non-weighted values are reported. The weighted means and standard deviations of the variables are estimated based on the log of leaving the school weights. In terms of the three aspects of principal leadership, principal leadership related to student management has the largest mean and standard

deviation followed by principal leadership related to creating a supportive school culture and instructional leadership. However, given the different items used for measuring those three aspects, such a simple comparison might not be meaningful. The means and standard deviations of the general leadership measures slightly decreased as ECTs gained experience from Year 3 to Year 5. In terms of school and teacher characteristics (weighted), 31% of the sampled teachers were male and 92% of the teachers were white; 61% of the teachers were school union members, and 20% of the teachers had a master's degree or higher. Only 5% of the teachers worked at charter schools in Year 1, and 64% of the schools made AYP in the 2006-07 school year; 26% of the teachers worked at schools located in suburban areas and 30% of the teachers worked at elementary schools. Whereas it is hard to detect any patterns between weighted means and non-weighted means, weighted standard deviations tended to be smaller than non-weighted ones. That is, applying weights reduces variation within each variable, so using the weights might reduce the statistical power of the analysis.

Table 20. Descriptive statistics

	Mean		Standard deviation	
	Weighted	Non-weighted	Weighted	Non-weighted
<i>Principal leadership</i>				
Instructional leadership (year 1)	1.647	1.678	0.354	0.446
Leadership related to student management (year 1)	3.023	2.991	0.643	0.663
Leadership related to supportive culture (year 1)	2.951	2.937	0.476	0.508
General leadership (year 1)	3.012	2.678	0.525	0.529
General leadership (year 3)	3.515	3.521	0.641	0.635
General leadership (year 4)	3.503	3.519	0.634	0.623
General leadership (year 5)	3.478	3.488	0.63	0.628
<i>Teacher Characteristics</i>				
Preparation	2.987	2.975	0.552	0.558
Alternative certified	0.267	0.275	0.442	0.447
Induction	0.755	0.716	0.43	0.451
Work hours per week	54.305	53.909	10.616	11.112
Autonomy	3.274	3.282	0.5	0.511
Male	0.314	0.321	-	-
White	0.92	0.878	-	-
Union membership	0.605	0.576	-	-
Advanced degree	0.199	0.201	-	-
Salary(logged) (year 1)	10.407	10.394	0.248	0.276
Salary(logged) (year 2)	10.528	10.463	0.184	0.233
Salary(logged) (year 3)	10.564	10.493	0.18	0.229

Table 20 (cont'd)

	Mean		Standard deviation	
	Weighted	Non-weighted	Weighted	Non-weighted
Salary(logged) (year 4)	10.605	10.528	0.212	0.277
Salary(logged) (year 5)	10.613	10.57	0.215	0.233
Highly qualified teacher (year 1)	0.703	0.661	-	-
Highly qualified teacher (year 2)	0.883	0.855	-	-
Highly qualified teacher (year 3)	0.881	0.869	-	-
Highly qualified teacher (year 4)	0.93	0.912	-	-
Highly qualified teacher (year 5)	0.963	0.958	-	-
Teaching demanding subject	0.312	0.313	-	-
Commitment	2.315	2.306	0.568	0.588
<i>School Characteristics</i>				
Percentage of FRL eligible students (year 1)	44.185	45.276	26.649	27.321
Percentage of FRL eligible students (year 2)	46.269	45.186	27.693	25.407
Percentage of FRL eligible students (year 3)	45.124	44.515	27.776	25.115
Percentage of FRL eligible students (year 4)	46.74	46.293	27.578	25.011
Percentage of FRL eligible students (year 5)	45.966	45.431	27.57	25.148
Charter school (year 1)	0.046	0.064	-	-
Charter school (year 2)	0.043	0.053	-	-
Charter school (year 3)	0.036	0.049	-	-
Charter school (year 4)	0.038	0.05	-	-
Charter school (year 5)	0.037	0.047	-	-
School size	833.213	813.884	618.979	650.353
Percentage of racially minority students	40.047	43.227	33.473	35.41
School safety	4.071	4.078	0.424	0.433
Parents' involvement	2.455	2.427	0.775	0.776
AYP status	0.638	0.638	-	-
Suburban (year 1)	0.255	0.234	-	-
Suburban (year 2)	0.305	0.25	-	-
Suburban (year 3)	0.324	0.254	-	-
Suburban (year 4)	0.322	0.256	-	-
Suburban (year 5)	0.327	0.266	-	-
Elementary school	0.295	0.247	-	-
Middle school	0.239	0.266	-	-
High school	0.313	0.487	-	-
<i>Weight</i>				
Leaving the school weight	-	85.845	-	153.145
Leaving the school weight(logged)	-	3.312	-	1.836
Leaving the profession weight	-	91.536	-	161.295
Leaver the profession weight(logged)	-	3.393	-	1.829

Note. All weighted mean and standard deviation were calculated based on log of leaving the school weight. The variables male, white, union membership, advanced degree, HQT, teaching demanding

Table 20 (cont'd)

subject (i.e., mathematics, science, ESL, and special education), charter school, AYP status, suburban, elementary school, and high school are dichotomous variables. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were the only variables for which data were available each year (i.e., time-variant variables). The combined school category (i.e., including at least one grade lower than 7 and at least one grade higher than 8) is included with the middle school category.

Table 21 reports the results from the survivor functions. With regard to the survivor function for teachers leaving their schools, in the second year, 74% of teachers worked at their original schools, and about a half of the cohort worked at their original schools until their fifth year. Based on the survivor function for the leaving the profession, in the second year, 90% of teachers stayed in the teaching profession, and 76% of the cohort were still working as teachers in their fifth year.

Table 21. Survivor Function

	Leaving the school	Leaving the profession
2 nd year	0.7434	0.9006
3 rd year	0.6308	0.8481
4 th year	0.5501	0.7972
5 th year	0.5015	0.7635

Note. The survivor function was weighted by log of weights.

Figures 7 and 8 show Kaplan-Meier survival curves by each principal leadership quartile. In terms of leaving the school, principal leadership related to creating a supportive culture showed a clear match between the turnover level and the principal leadership quartile; ECTs who perceived their principal as a strong leader in their first year left their schools at lower rates throughout the first five years of their teaching career. With regard to the other two aspects of leadership, the order of the survival rates and the leadership quartiles were not as clearly matched. For example, ECTs in the highest quartile of instructional leadership and leadership related to student management did not have the lowest rates in terms of leaving their schools. However, the lowest quartiles in those leadership scales still showed the lowest survival rates. The survival curves for leaving the profession did not show a clear association between principal leadership quartiles and ECTs' survival rates. The only consistent pattern is that the lowest quartiles for all three aspects showed the lowest survival rates throughout the five years. This implies a weaker association between principal leadership and teachers leaving the profession.

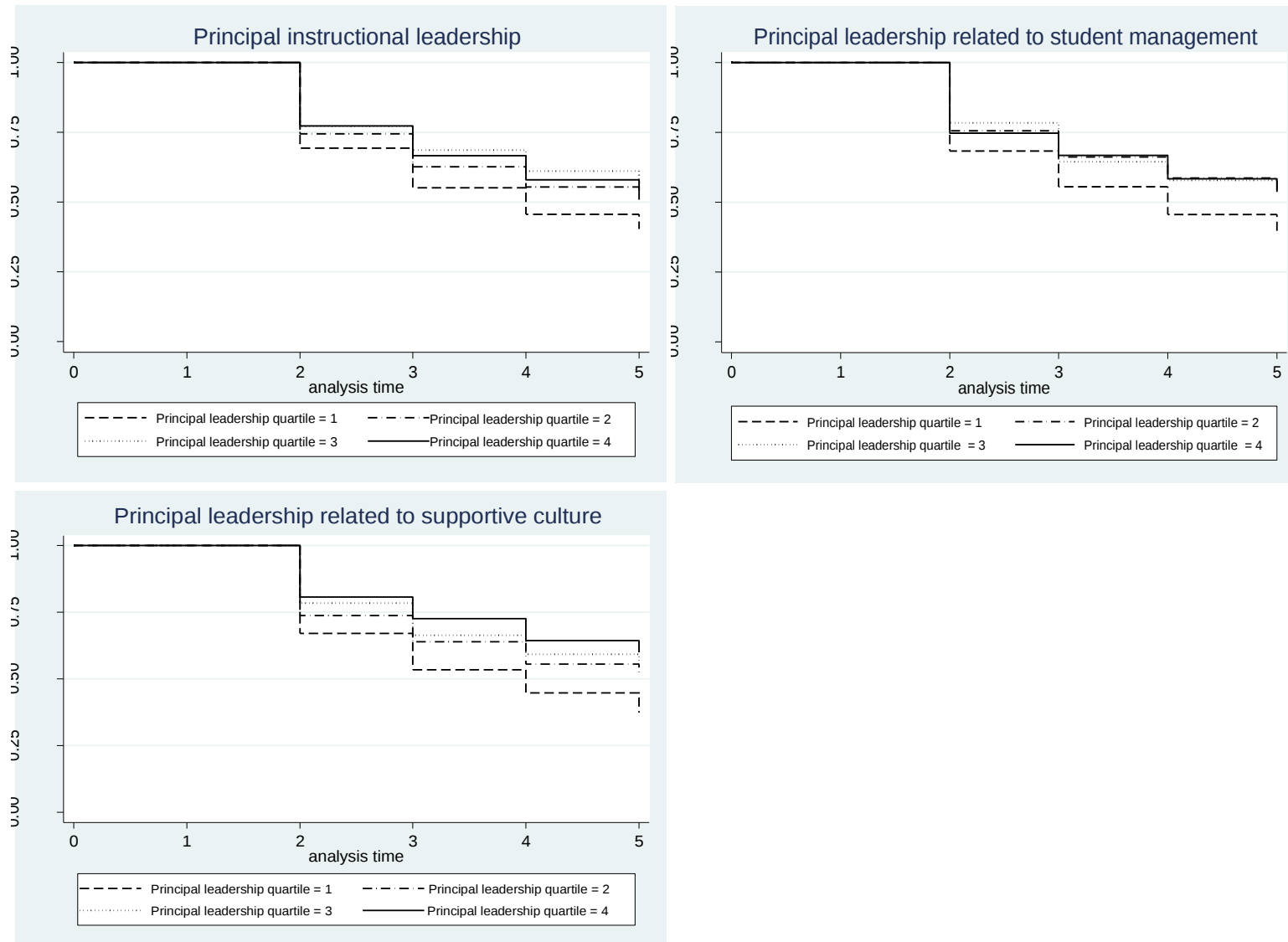


Figure 7. Kaplan-Meier survival Curves for Teachers Leaving the School

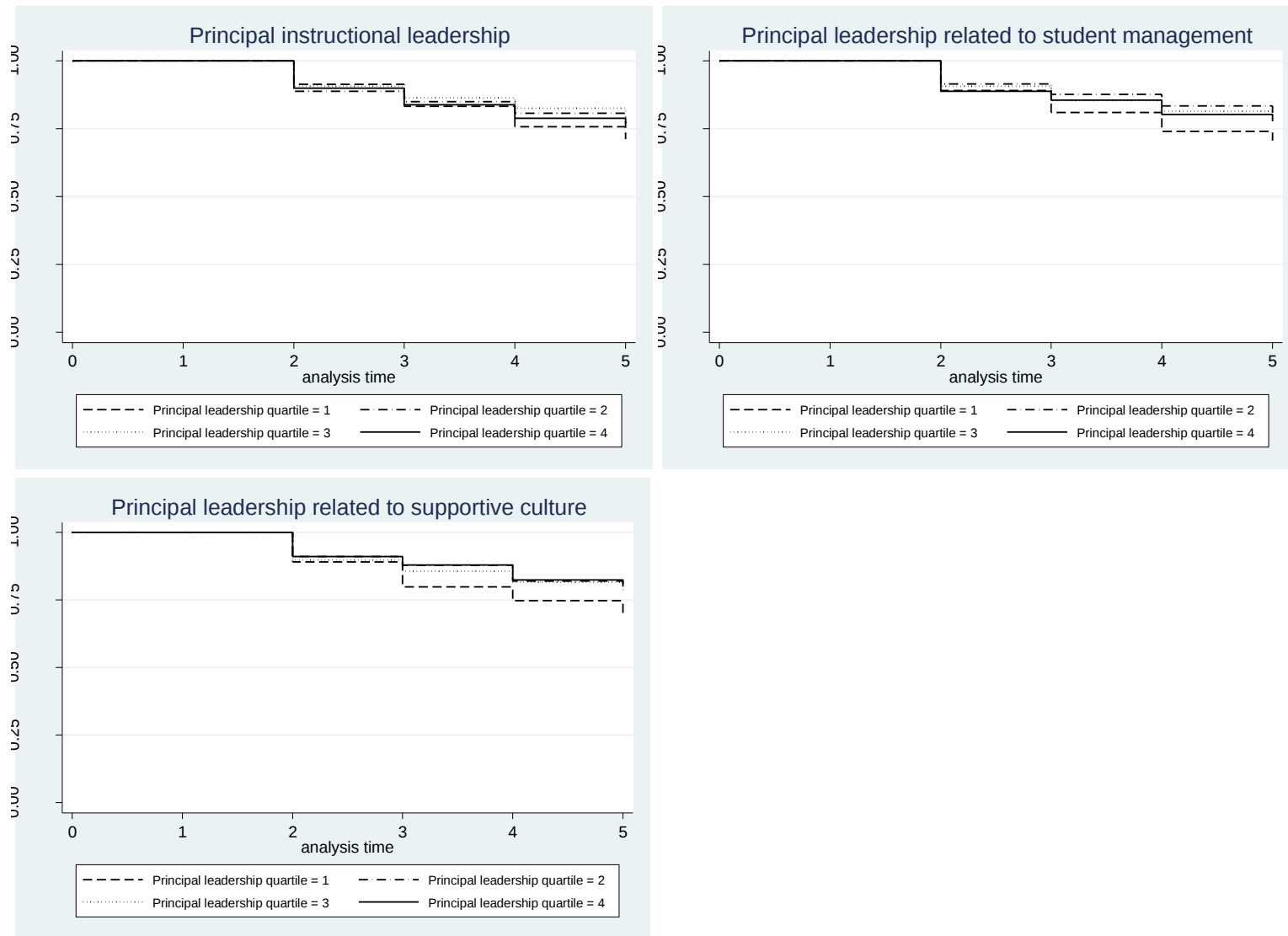


Figure 8. Kaplan-Meier Survival Curves for Teacher Leaving the Profession

I also conducted a univariate Cox proportional hazard regression to test whether ECTs in different principal leadership quartiles had a significantly different survival function in the five years. Table 22 reports the results. Based on the Wald-test results, it was possible to reject the null hypothesis that the survival curves of each leadership quartile are all the same for the outcomes (i.e., leaving the school and leaving the profession). That is, it is valid to include the principal leadership measures to predict the likelihood of ECT turnover.

Table 22. Univariate Cox Regression Based Test for Quality of Survival Curves

<i>Panel A. Analysis on leaving the school</i>									
	Instructional leadership			Leadership related to student management			Leadership related to supportive culture		
	Events observed (EO)	Events expected (EP)	Relative hazard (RH)	EO	EP	RH	EO	EP	RH
1 st	967.01	765.30	1.2790	1049.34	823.95	1.29	1235.79	919.49	1.38
2 nd	1075.88	1095.88	0.9934	793.94	869.70	0.92	719.25	757.71	0.97
3 rd	840.37	1006.92	0.8443	836.25	927.07	0.91	831.43	949.91	0.9
4 th	398.57	413.72	0.9747	597.95	656.76	0.92	495.34	654.71	0.77
Wald test		24.28***			25.27***			48.65***	
(p-value)		(<0.0001)			(<0.0001)			(<0.0001)	
<i>Panel B. Analysis on leaving the profession</i>									
	Instructional leadership			Leadership related to student management			Leadership related to supportive culture		
	EO	EP	RH	EO	EP	RH	EO	EP	RH
1 st	478.34	388.82	1.2438	529.44	417.78	1.28	614.19	466.41	1.34
2 nd	517.20	530.31	0.9860	360.25	421.41	0.87	316.62	374.19	0.86
3 rd	391.55	474.34	0.8345	407.58	439.36	0.94	406.74	446.83	0.93
4 th	206.00	199.62	1.0433	295.82	314.54	0.95	255.54	305.66	0.85
Wald test		9.2*			10.98*			17.26***	
(p-value)		(0.0268)			(0.0118)			(0.0006)	

Note. The survivor function was weighted by log of weights. * p<0.05, ** p<0.01, *** p<0.001

Although the results from the univariate analysis suggested a significant relationship between principal leadership and ECTs' career trajectory, it is important to include other control variables to verify such an association, and to estimate the magnitude of the association. Moreover, dividing ECTs into quartiles based on their perceptions about their principal leadership involves an arbitrary decision regarding the criteria for dividing groups. In addition, with this framework, it is hard to incorporate time-

variant principal leadership variables. Thus, building on the exploratory results from the univariate analysis, I proceed with a discrete-time survival analysis of teacher turnover.

Discrete time survival analysis on ECTs leaving the school

Table 23 represents the results for teachers leaving their schools using a log of weights. Models 1 through 4 did not include any teacher-level or school-level control variables; Models 6 through 10 included teacher background variables; Models 11 through 16 included school characteristics and teacher backgrounds as control variables. For ease of interpretation, the coefficients are reported as odds ratios. Without any control variables, all three aspects of principal leadership measured at the first year and general principal leadership measured during the previous year had a significant association with teachers leaving the school during the first five years of ECTs' careers. A one-unit increase in principals' general leadership in the previous year was associated with a 41% lower odds of teachers leaving the school in any given year during the first five years, after controlling for their base line survival function (odds ratio=0.593, $p \leq 0.001$). Two-year lagged general principal leadership did not have a significant relationship with the outcome after controlling for one-year lagged principal leadership. Among the three aspects, principal leadership related to creating a supportive culture had the strongest association with the odds of teachers leaving the school. A one-unit increase in principal leadership related to creating a supportive school culture in the first year of ECTs' teaching experience was associated with a 45% lower odds of teachers leaving the school in any given year during the five years (odds ratio=0.551, $p \leq 0.001$). This was slightly larger than the effects of general principal leadership. The other two aspects of principal leadership, instructional leadership and leadership related to student management, also had negative and significant associations with ECTs leaving the school (odds ratio=0.724, $p \leq 0.001$; odds ratio=0.758, $p \leq 0.001$, respectively).

Table 23. The Influence of Principal Leadership on Leaving the School

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.593*** (0.0375)	0.663*** (0.0563)				0.655*** (0.052)	0.708*** (0.0686)	
General leadership (2 year lagged)		0.846 (0.0744)					0.838 (0.0789)	
Instructional leadership			0.724*** (0.0727)					0.792 (0.113)
Leadership related to student management				0.758*** (0.0467)				
Leadership related to supportive culture					0.551*** (0.0438)			
Preparation						0.892 (0.0753)	0.852 (0.0798)	0.898 (0.0761)
Alternative certificate						0.934 (0.1)	0.961 (0.113)	0.945 (0.101)
Induction						0.871 (0.0894)	0.867 (0.0998)	0.867 (0.0893)
Work hours						0.992 (0.0042)	0.996 (0.0048)	0.991* (0.00428)
Autonomy						0.846 (0.0779)	0.872 (0.0903)	0.805 (0.0743)
Male						0.94 (0.0883)	1.041 (0.111)	0.919 (0.0866)
White						0.689* (0.109)	0.693* (0.124)	0.688* (0.106)
Union membership						0.944 (0.0876)	0.945 (0.0986)	0.965 (0.0895)
Advanced Degree						0.992 (0.116)	1.047 (0.135)	0.986 (0.115)
Salary(log)						0.618** (0.113)	0.611** (0.117)	0.632* (0.115)
HQT						0.954 (0.103)	1.014 (0.128)	0.951 (0.102)
PD						1.041 (0.0228)	1.047 (0.0253)	1.039 (0.0227)

Table 23 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Demanding subject						1.022 (0.0989)	1.053 (0.113)	1.034 (0.0999)
Commitment						0.626*** (0.053)	0.736*** (0.0685)	0.585*** (0.0501)
N	4400	3750	4400	4410	4410	3960	3370	3970
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
General leadership (1 year lagged)			0.7*** (0.0637)	0.766* (0.0861)				
General leadership (2 year lagged)				0.87 (0.0939)				
Instructional leadership					0.813 (0.126)			1.049 (0.191)
Leadership related to student management	0.873 (0.0656)					0.885 (0.0786)		1.023 (0.1054)
Leadership related to supportive culture		0.633*** (0.0620)					0.680*** (0.0768)	0.656*** (0.0912)
Preparation	0.892 (0.0754)	0.896 (0.0757)	0.853 (0.0781)	0.791* (0.0801)	0.857 (0.0787)	0.854 (0.0781)	0.852 (0.0777)	0.850 (0.078)
Alternative certificate	0.937 (0.101)	0.929 (0.1)	0.950 (0.113)	0.968 (0.126)	0.948 (0.112)	0.952 (0.112)	0.951 (0.113)	0.952 (0.113)
Induction	0.862 (0.0885)	0.888 (0.092)	0.919 (0.103)	0.906 (0.113)	0.919 (0.103)	0.914 (0.103)	0.93 (0.105)	0.93 (0.105)
Work hours	0.992 (0.00422)	0.992 (0.00425)	0.990* (0.00453)	0.994 (0.00516)	0.990* (0.00462)	0.990* (0.00456)	0.990* (0.00459)	0.990* (0.005)
Autonomy	0.805* (0.0743)	0.847 (0.0787)	0.869 (0.0901)	0.852 (0.101)	0.834 (0.0866)	0.834 (0.0864)	0.872 (0.0914)	0.871 (0.0914)
Male	0.9 (0.0850)	0.929 (0.0872)	0.878 (0.0915)	0.997 (0.116)	0.862 (0.0899)	0.85 (0.0886)	0.875 (0.0909)	0.876 (0.0914)
White	0.699* (0.109)	0.699* (0.109)	0.731 (0.127)	0.74 (0.147)	0.743 (0.126)	0.746 (0.13)	0.734 (0.125)	0.737 (0.126)
Union membership	0.965 (0.0892)	0.945 (0.0879)	0.964 (0.0991)	0.955 (0.111)	0.984 (0.101)	0.983 (0.101)	0.962 (0.0992)	0.963 (0.0993)

Table 23 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Advanced Degree	1.007 (0.118)	0.993 (0.117)	1.027 (0.131)	1.081 (0.152)	1.026 (0.131)	1.044 (0.133)	1.037 (0.133)	1.043 (0.134)
Salary(log)	0.632* (0.115)	0.609** (0.111)	0.532*** (0.1)	0.539** (0.107)	0.535*** (0.101)	0.542** (0.102)	0.522*** (0.0982)	0.524*** (0.099)
HQT	0.959 (0.103)	0.945 (0.102)	0.937 (0.111)	0.97 (0.133)	0.934 (0.11)	0.939 (0.111)	0.926 (0.109)	0.924 (0.109)
PD	1.041 (0.0227)	1.046* (0.0229)	1.038 (0.0249)	1.043 (0.0274)	1.036 (0.0248)	1.037 (0.0248)	1.042 (0.0251)	1.041 (0.0251)
Demanding subject	1.021 (0.0985)	1.022 (0.0987)	1.031 (0.110)	1.076 (0.127)	1.045 (0.111)	1.039 (0.11)	1.034 (0.109)	1.031 (0.1009)
Commitment	0.589*** (0.0504)	0.639*** (0.0552)	0.603*** (0.0556)	0.708*** (0.0724)	0.573*** (0.0533)	0.576*** (0.0535)	0.613*** (0.0577)	0.610*** (0.058)
School size			1 (0.0001)	1 (0.00011)	1 (0.0001)	1 (0.0001)	1 (0.0001)	1 (0.0001)
% of racially minority students			1.003 (0.002)	1.004 (0.00221)	1.004 (0.00198)	1.004 (0.00199)	1.003 (0.00201)	1.003 (0.00201)
School safety			0.935 (0.118)	0.906 (0.129)	0.919 (0.116)	0.94 (0.12)	0.946 (0.12)	0.945 (0.121)
Parents' involvement			0.924 (0.072)	0.954 (0.0864)	0.913 (0.0713)	0.921 (0.0717)	0.92 (0.0719)	0.919 (0.072)
AYP status			0.762* (0.0819)	0.747* (0.0894)	0.758** (0.081)	0.757** (0.0814)	0.776* (0.0835)	0.774* (0.084)
% of FRL			1 (0.00278)	0.999 (0.00315)	1 (0.00276)	1 (0.00276)	1 (0.00278)	1 (0.00278)
Suburban			0.937 (0.116)	0.868 (0.119)	0.944 (0.117)	0.94 (0.117)	0.969 (0.12)	0.971 (0.12)
Charter school			0.706 (0.171)	0.58 (0.166)	0.714 (0.173)	0.704 (0.171)	0.698 (0.17)	0.967 (0.17)
Elementary school			1.139 (0.166)	1.055 (0.171)	1.139 (0.166)	1.153 (0.168)	1.144 (0.167)	1.142 (0.167)
High school			0.881 (0.117)	0.822 (0.121)	0.846 (0.112)	0.843 (0.112)	0.864 (0.115)	0.867 (0.116)

Table 23 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
N	3960	3970	3510	2980	3510	3500	3510	3500

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). Estimates are adjusted by log of leaving the school weights. The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

After controlling for school and teacher backgrounds (Models 11 through 15), only time-variant principal general leadership and principal leadership related to creating a supportive culture had a significant association with the odds of teachers leaving the school. Although the magnitude of the associations was mitigated and the standard errors of the main variables were increased by including the control variables, the coefficients of these two principal leadership variables remained substantial (odds ratios=0.7 and 0.68, respectively, after including all control variables). The other two aspects of principal leadership became statistically insignificant with lower coefficients and higher standard errors. Besides principal leadership, teachers' race, work hours, salary, school AYP status, and teacher commitment had a significant association with the outcome in some models; among these variables, salary and commitment had a strong negative association even after including school characteristics in the models. For those models, student composition itself (i.e., student race and SES levels) did not have any association with teacher turnover.

Model 16 includes all three aspects of principal leadership at the same time in order to compare the effects of each aspect of principal leadership. While the other two aspects of leadership were not significant and their odds ratios were close to 1, principal leadership related to creating a supportive culture remained significant and its coefficient stayed at a similar level as the previous models (odds ratio=0.656, $p \leq 0.01$). However, given a considerable increase in standard errors of each principal leadership variable, this result might be less reliable, due to potential multicollinearity. Accordingly, I used the Hausman test to see whether the coefficients for each principal leadership variable in models 13 through 15 were statistically different from each other. It should be noted that the Hausman test is generally less conservative compared to other methods, such as the Cohen and Cohen test or the Wald test (Cohen & Cohen, 1983). As a result, the impact of principal leadership related to creating a supportive culture is statistically different from the impact of instructional leadership at the level of $\alpha=0.1$ ($\chi^2=3.14$, $p=0.076$), and it was statistically different from the impact of principal leadership related to student management at the level of $\alpha=0.0001$ ($\chi^2=15.038$, $p=0.000105$). Thus, it is valid to argue that principal

leadership related to creating a supportive culture had a stronger association with ECTs leaving the school, compared to other the two aspects of principal leadership.

Table 24. The Heterogeneous Effects of Principal Leadership on Leaving the School

<i>Panel A. Principal leadership*time indicators</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.759* (0.0849)			
Instructional leadership		0.768 (0.168)		
Leadership related to student management			0.845 (0.098)	
Leadership related to supportive culture				0.620* (0.1)
Year 2	15259.2*** (30932.7)	17345.5*** (35399.1)	14345.0*** (29072.6)	38653.3*** (79601.1)
Year 3	10465.9*** (21312.2)	10248.6*** (21009.9)	8470.5*** (17258.9)	23154.7*** (47947.7)
Year 4	11669.3*** (24007.6)	9278.4*** (19135)	7653.8*** (15682)	20830.8*** (43358.4)
Year 5	8046.9*** (16605.8)	6787.5*** (13948.5)	5646.4*** (11541)	15298.8*** (31782.9)
Year 2* Leadership	0.669 (0.173)	1.221 (0.602)	1.162 (0.282)	1.309 (0.482)
Year 3* Leadership	0.91 (0.291)	1.196 (0.674)	1.187 (0.315)	1.682 (0.682)
Year 4* Leadership	0.799 (0.223)	1.244 (0.71)	1.133 (0.321)	1.221 (0.527)
N	3510	3510	3500	3510
<i>Panel B. Principal leadership*Suburban</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.7*** (0.0637)			
Instructional leadership		0.810 (0.124)		
Leadership related to student management			0.885 (0.0786)	
Leadership related to supportive culture				0.666*** (0.0747)
Suburban	0.939 (0.124)	0.934 (0.117)	0.939 (0.117)	0.963 (0.120)
Suburban*Leadership	1.010 (0.186)	0.768 (0.235)	0.963 (0.174)	0.602* (0.140)
N	3510	3510	3500	3510

Table 24 (cont'd)

*Panel C. Principal leadership*Elementary school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.689*** (0.0626)			
Instructional leadership		0.798 (0.124)		
Leadership related to student management			0.876 (0.0771)	
Leadership related to supportive culture				0.649*** (0.0740)
Elementary school	1.203 (0.182)	1.147 (0.167)	1.099 (0.165)	1.152 (0.168)
Elementary school*Leadership	1.237 (0.204)	1.205 (0.348)	1.351 (0.244)	1.634* (0.344)
N	3510	3510	3500	3510

*Panel D. Principal leadership*High school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.703*** (0.0649)			
Instructional leadership		0.818 (0.127)		
Leadership related to student management			0.884 (0.0781)	
Leadership related to supportive culture				0.660*** (0.0747)
High school	0.89 (0.123)	0.849 (0.114)	0.843 (0.112)	0.850 (0.114)
High school* Leadership	1.042 (0.158)	1.063 (0.284)	0.992 (0.151)	0.747 (0.147)
N	3510	3510	3500	3510

*Panel E. Principal leadership*Charter school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.701*** (0.0637)			
Instructional leadership		0.815 (0.126)		
Leadership related to student management			0.887 (0.079)	
Leadership related to supportive culture				0.679*** (0.0767)
Charter school	0.669 (0.180)	0.681 (0.173)	0.665 (0.168)	0.71 (0.175)
Charter school* Leadership	0.860 (0.267)	0.630 (0.339)	0.689 (0.222)	1.118 (0.46)
N	3510	3510	3500	3510

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The variables for calculating interaction terms were centered by their grand means. Estimates

Table 24 (cont'd)

are adjusted by log of leaving the school weights. The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In terms of heterogeneous effects of principal leadership (i.e., research questions 2 and 3), I included four different sorts of interaction terms [See Table 24]. As noted earlier, the variables used for calculating the interaction terms were grand-mean centered, in order to avoid multicollinearity. The first panel presents the results using terms for the interactions between time indicators and principal leadership variables; none of the interaction terms were significant. That is, the effects of principal leadership on the outcome variable were relatively constant across five years. This result justified the use of a discrete time survival analysis, which has only one coefficient for each independent variable across years. The second set of models included terms for the interactions between an indicator of whether a school was located in a suburban area and the principal leadership variables (*Panel B*). Only a term for the interaction between principal leadership related to creating a supportive culture and suburban location was significant in Model 4. The association between principal leadership related to creating a supportive culture and the odds of teachers leaving the school was stronger when a teacher worked at a school located in the suburbs. In terms of the effects of principal leadership at elementary schools, only a term for the interaction between an indicator of whether a teacher worked at an elementary school and principal leadership related to creating a supportive culture had a significant association with teacher turnover (*Panel C*). For elementary school teachers, principal leadership related to creating a supportive culture was less influential for the odds of leaving their school when compared to teachers who worked at other school levels. Other interaction terms related to the elementary school indicator were not significant. None of terms for the interactions between an indicator of working at a high school or working at a charter schools and principal leadership variables were significant (*Panel D and E*).

Using raw turnover weights (i.e., weights that have not been transformed) did not lead to substantial changes in the main results, except that principal leadership related to student management had a significant negative association with the odds of teachers leaving the school along with time-variant

general principal leadership and principal leadership related to creating a supportive culture [See Tables 34 and 35 in Appendix F]. The magnitude of the associations between the principal leadership measures and the outcome was similar to that in models using the log of weights. Instructional leadership was not significant in any models, and none of the interaction terms were significant. The same analysis without any weights produced similar results [See Appendix G].

Discrete time survival analysis on teachers leaving the profession

Table 25 presents the results from a discrete time survival analysis for teachers leaving the profession. I applied log of leave-the-profession weights to the analysis. Without any control variables, all three aspects of principal leadership and time-variant general principal leadership variables had a significant negative association with the likelihood of teachers leaving the profession. Again, the two-year lagged general principal leadership variable had no significant association with the outcome in any models. After including teacher- and school-level control variables, only general principal leadership had a significant effect on teachers leaving the profession. A one-unit increase in principal leadership in the previous year was associated with a 27% lower odds of teachers leaving the profession. Interestingly, the magnitude of this variable was much smaller than the one from the analysis for teachers leaving the school. As including teacher- and school-level control variables reduced the coefficients and increased the standard errors of the other principal leadership variables, none of the three aspects of principal leadership had a significant association with teachers leaving the profession. Instead, the magnitude of the coefficients for salary, commitment, and preparation increased; in particular, the association between salary and teachers leaving the profession was substantial (odds ratio=0.361, $p \leq 0.001$), while school AYP status was no longer significant. This suggests that the attributes of the teaching profession as a whole, such as salaries, and characteristics of individual teachers themselves (e.g., their commitment and perceptions about their preparation) are more influential for teachers leaving the profession, than school organizational factors, such as principal leadership or AYP status. Since none of the three aspects of principal leadership were significant, I did not compare the coefficients for those variables with additional tests, such as the Hausman test.

Table 25. The Influence of Principal Leadership on Leaving the Profession

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.661*** (0.0563)	0.668*** (0.0725)				0.723** (0.073)	0.712** (0.0875)	
General leadership (2 year lagged)		0.903 (0.108)					0.905 (0.116)	
Instructional leadership			0.738* (0.104)					0.731 (0.14)
Leadership related to student management				0.667*** (0.0687)				
Leadership related to supportive culture					0.764** (0.065)			
Preparation						0.749** (0.0797)	0.739** (0.0853)	0.762* (0.0813)
Alternative certificate						0.928 (0.129)	0.968 (0.148)	0.949 (0.131)
Induction						0.723* (0.0939)	0.773 (0.111)	0.722* (0.0939)
Work hours						0.992 (0.00539)	0.994 (0.0061)	0.991 (0.00554)
Autonomy						0.991 (0.118)	1.011 (0.13)	0.969 (0.115)
Male						1.061 (0.136)	1.059 (0.149)	1.039 (0.133)
White						0.792 (0.157)	0.911 (0.21)	0.784 (0.153)
Union membership						0.953 (0.117)	0.943 (0.126)	0.965 (0.118)
Advanced Degree						1.125 (0.174)	1.126 (0.19)	1.102 (0.17)
Salary(log)						0.383*** (0.091)	0.373*** (0.0941)	0.387*** (0.0899)
HQT						0.778 (0.11)	0.911 (0.155)	0.776 (0.11)
PD						1.041 (0.0295)	1.035 (0.0322)	1.04 (0.0293)

Table 25 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Demanding subject						1.158 (0.145)	1.22 (0.167)	1.163 (0.145)
Commitment						0.633*** (0.0655)	0.712** (0.0804)	0.614*** (0.065)
N	5290	4630	5320	5320	5300	4770	4170	4780
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
General leadership (1 year lagged)			0.731** (0.0865)	0.699* (0.106)				
General leadership (2 year lagged)				0.985 (0.155)				
Instructional leadership					0.784 (0.174)			0.847 (0.214)
Leadership related to student management	0.885 (0.0893)					0.871 (0.104)		0.966 (0.187)
Leadership related to supportive culture		0.833 (0.114)					0.857 (0.137)	0.966 (0.1876)
Preparation	0.753** (0.0803)	0.751** (0.08)	0.676** (0.0816)	0.656** (0.0857)	0.682** (0.0826)	0.680** (0.0822)	0.675** (0.0813)	0.684*** (0.083)
Alternative certificate	0.945 (0.131)	0.947 (0.131)	0.825 (0.137)	0.844 (0.153)	0.825 (0.136)	0.828 (0.137)	0.829 (0.137)	0.826 (0.136)
Induction	0.714** (0.0924)	0.719* (0.0934)	0.823 (0.124)	0.860 (0.143)	0.820 (0.123)	0.814 (0.122)	0.816 (0.123)	0.821 (0.124)
Work hours	0.992 (0.00545)	0.992 (0.0054)	0.989 (0.00584)	0.991 (0.00665)	0.988* (0.00605)	0.989 (0.00596)	0.989 (0.00587)	0.988* (0.006)
Autonomy	0.962 (0.114)	0.97 (0.116)	0.958 (0.134)	0.942 (0.145)	0.928 (0.130)	0.927 (0.129)	0.931 (0.131)	0.940 (0.133)
Male	1.021 (0.131)	1.039 (0.133)	0.980 (0.146)	0.961 (0.158)	0.959 (0.142)	0.945 (0.140)	0.958 (0.142)	0.953 (0.142)
White	0.792 (0.155)	0.793 (0.156)	0.747 (0.166)	0.833 (0.212)	0.752 (0.166)	0.753 (0.167)	0.757 (0.167)	0.747 (0.165)
Union membership	0.968 (0.118)	0.963 (0.118)	0.973 (0.139)	0.932 (0.148)	1.002 (0.143)	1.001 (0.143)	0.998 (0.143)	0.996 (0.142)
Advanced Degree	1.127 (0.175)	1.124 (0.173)	1.136 (0.200)	1.189 (0.231)	1.125 (0.197)	1.143 (0.201)	1.144 (0.201)	1.128 (0.198)

Table 25 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Salary(log)	0.389*** (0.0909)	0.39*** (0.091)	0.356*** (0.0815)	0.343*** (0.0877)	0.359*** (0.0810)	0.361*** (0.0819)	0.361*** (0.0818)	0.358*** (0.081)
HQT	0.782 (0.111)	0.779 (0.11)	0.775 (0.123)	0.880 (0.167)	0.775 (0.123)	0.776 (0.123)	0.774 (0.123)	0.772 (0.123)
PD	1.04 (0.0294)	1.04 (0.0292)	1.053 (0.0333)	1.047 (0.0354)	1.050 (0.0329)	1.052 (0.0330)	1.051 (0.0328)	1.052 (0.033)
Demanding subject	1.148 (0.144)	1.156 (0.144)	1.109 (0.157)	1.165 (0.182)	1.116 (0.157)	1.104 (0.156)	1.107 (0.156)	1.110 (0.157)
Commitment	0.606*** (0.0635)	0.614*** (0.067)	0.590*** (0.0712)	0.674** (0.0924)	0.574*** (0.0705)	0.573*** (0.0695)	0.573*** (0.0719)	0.587 (0.075)***
School size			1.000 (0.000142)	1.000 (0.000162)	1.000 (0.000142)	1.000 (0.000143)	1.000 (0.000143)	1.000 (0.000142)
% of racially minority students			1.003 (0.00268)	1.004 (0.00284)	1.003 (0.00264)	1.003 (0.00267)	1.003 (0.00267)	1.003 (0.00267)
School safety			1.118 (0.188)	1.214 (0.225)	1.072 (0.181)	1.093 (0.185)	1.079 (0.182)	1.091 (0.184)
Parents' involvement			0.914 (0.0999)	0.940 (0.117)	0.911 (0.0992)	0.918 (0.100)	0.910 (0.0993)	0.917 (0.1004)
AYP status			0.861 (0.126)	0.895 (0.144)	0.866 (0.127)	0.865 (0.126)	0.865 (0.126)	0.875 (0.129)
% of FRL			0.997 (0.00350)	0.996 (0.00386)	0.997 (0.00350)	0.996 (0.00348)	0.997 (0.00351)	0.996 (0.003)
Suburban			0.855 (0.153)	0.815 (0.162)	0.858 (0.152)	0.852 (0.151)	0.860 (0.153)	0.858 (0.152)
Charter school			0.980 (0.303)	0.751 (0.277)	1.034 (0.314)	1.020 (0.308)	1.024 (0.311)	1.023 (0.310)
Elementary school			0.941 (0.192)	0.771 (0.175)	0.939 (0.191)	0.949 (0.192)	0.940 (0.191)	0.948 (0.192)
High school			1.132 (0.213)	1.112 (0.227)	1.089 (0.204)	1.086 (0.204)	1.100 (0.206)	1.087 (0.204)
N	4770	4780	4060	3540	4070	4070	4060	4060

Table 25 (cont'd)

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). Estimates are adjusted by log of leaving the profession weights. The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 26. The Heterogeneous Effects of Principal Leadership on Leaving the Profession

<i>Panel A. Principal leadership*time indicators</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.716* (0.0949)			
Instructional leadership		0.749 (0.202)		
Leadership related to student management			0.881 (0.126)	
Leadership related to supportive culture				0.845 (0.167)
Year 2	417669.1*** (1033030.0)	324018.5*** (780229.7)	260906.2*** (631712.1)	311608.2*** (757906.3)
Year 3	267574.7*** (668204.6)	185586.7*** (449215.6)	155859.9*** (379443.2)	190799.1*** (467130.4)
Year 4	447286.7*** (1122375.3)	264624.4*** (644777.2)	219154.2*** (537129.3)	262713.2*** (646871.6)
Year 5	349149.8*** (878900.2)	201770.8*** (488751.5)	165101.2*** (402998.5)	196310.1*** (481670.9)
Year 2* Leadership	1.1 (0.337)	1.238 (0.69)	0.965 (0.285)	1.056 (0.451)
Year 3* Leadership	1.169 (0.414)	0.574 (0.373)	0.704 (0.237)	0.892 (0.41)
Year 4* Leadership	1.139 (0.354)	0.9 (0.562)	1.008 (0.346)	0.19 (0.566)
N	4060	4070	4060	4070
<i>Panel B. Principal leadership*Suburban</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.734** (0.0862)			
Instructional leadership		0.785 (0.175)		
Leadership related to student management			0.871 (0.104)	
Leadership related to supportive culture				0.864 (0.138)
Suburban	0.895 (0.163)	0.860 (0.156)	0.850 (0.152)	0.867 (0.155)
Suburban*Leadership	1.231 (0.239)	1.026 (0.461)	0.976 (0.248)	1.250 (0.422)
N	4060	4070	4060	4070

Table 26 (cont'd)

*Panel C. Principal leadership*Elementary school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.733* (0.0866)			
Instructional leadership		0.774 (0.172)		
Leadership related to student management			0.865 (0.103)	
Leadership related to supportive culture				0.832 (0.134)
Elementary school	0.931 (0.196)	0.953 (0.194)	0.913 (0.190)	0.961 (0.197)
Elementary school*Leadership	0.96 (0.195)	1.235 (0.539)	1.484 (0.378)	1.922* (0.608)
N	4060	4070	4060	4070

*Panel D. Principal leadership*High school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.739* (0.0881)			
Instructional leadership		0.788 (0.177)		
Leadership related to student management			0.877 (0.104)	
Leadership related to supportive culture				0.852 (0.136)
High school	1.158 (0.222)	1.095 (0.207)	1.106 (0.206)	1.095 (0.204)
High school*Leadership	1.118 (0.208)	1.065 (0.421)	1.194 (0.253)	0.933 (0.267)
N	4060	4070	4060	4070

*Panel E. Principal leadership*Charter school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.729** (0.0865)			
Instructional leadership		0.778 (0.174)		
Leadership related to student management			0.875 (0.105)	
Leadership related to supportive culture				0.854 (0.136)
Charter school	1.024 (0.321)	1.074 (0.331)	0.981 (0.319)	1.053 (0.323)
Charter school* Leadership	1.185 (0.353)	1.425 (0.911)	0.836 (0.303)	1.199 (0.582)
N	4060	4070	4060	4070

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible

Table 26 (cont'd)

students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The variables for calculating interaction terms were centered by their grand means. Estimates are adjusted by log of leaving the profession weights. The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Only one interaction term that was included in the previous analysis, the interaction between an indicator of elementary schools and principal leadership related to creating a supportive culture was significant [See Table 26]. As in the analysis for leaving the school, the association between principal leadership related to creating a supportive culture and teachers leaving the profession was weaker for elementary school teachers.

With non-transformed weights [See Table 36], none of the principal leadership variables had a significant association with teachers leaving the profession after the control variables were included. Interestingly, even the general principal leadership variable became insignificant as the coefficient decreased while the standard error increased. Still, teacher salary, commitment, and preparation had significant negative associations with leaving the profession. For the first time, teaching in-demand subjects had a significant association with the outcome. Teaching those subjects increased the odds of teachers leaving the profession by 1.5 times, but this variable became insignificant when school-level control variables were included. This may suggest that teachers who taught those in-demand subjects tended to have more options other than teaching positions. No interaction terms were significant, either [See Table 37]. In contrast, without any weights, principal leadership related to creating a supportive culture, and principal leadership related to student management, along with time-variant general principal leadership, had a significant association with teachers leaving the profession after including all of the control variables, although the magnitude of the coefficient was smaller than the one in the main analysis [See Table 39]. In terms of interaction terms, as in the previous analysis, for elementary school teachers, principal leadership related to creating a supportive culture tended to have a weaker effect on their leaving the profession. Other than this interaction term, none of the other interaction terms were significant in any models [See Table 41].

Lastly, I quantified the robustness of the results reported above by using Robustness Indices (Frank, 2000; Frank et al., 2008; Frank et al., 2013). I applied the Robustness Indices for three main inferences: 1) the association between general leadership and ECTs leaving the school; 2) the association between leadership related to creating supportive culture and ECTs leaving the school; and 3) the association between general leadership and ECTs leaving the profession. The first and second inferences seem to be quite strong. To invalidate these results, 51% and 43% of the sample would need to be replaced with cases for which there was no effect of principal leadership on teacher turnover, respectively (Frank et al., 2013). The inference about principal leadership and ECTs leaving the profession, however, turned out to be a relatively weak inference; replacing only 25% of sample would invalidate the inference.

Discussion

Using nationally representative data, the Beginning Teacher Longitudinal Survey, I found that ECT turnover was indeed substantial. About 25% of teachers left their school after one year and only 50% of teachers stayed at the same school after four years. These values are similar to ones reported by Allensworth and colleagues (2009), based on data from Chicago. Building on prior studies, this study examined principal leadership as one factor that potentially reduces ECT turnover rates. Three related aspects of principal leadership — instructional leadership, leadership related to student management, and leadership related to creating a supportive culture — along with general leadership were analyzed in light of their potential impact on ECTs' rates of leaving their schools and leaving the profession. Consistent with many previous research studies (Allensworth et al., 2009; Boyd et al., 2011; Grissom, 2011; Ingersoll & May, 2012; Ladd, 2011), principal leadership has a significant negative association with ECT turnover. However, this study expands on the previous studies in several ways.

First, I drew on five years of longitudinal data and showed that principal leadership has a constant and strong influence on ECTs leaving the school whereas previous studies typically focused on one or two years of data. In this study, whether it was measured in the first year or in the previous year, principal leadership had a significant association with teacher turnover, even after controlling for various school and teacher characteristics including student composition, teachers' salaries, and commitment levels, in

particular. Moreover, terms for the interactions between time indicators and principal leadership measures were never significant in any models. Contrary to the expectation that the influence of principal leadership would decrease as teachers gain experience, principal leadership was important for teacher turnover throughout ECTs' first five years in the profession. Although this was suggested by previous research studies that focused on teachers with various levels of experience (Allensworth et al., 2009; Boyd et al., 2011; Grissom, 2011; Ingersoll & May, 2012; Ladd, 2011), I showed the constant and strong influence of principal leadership on teacher turnover with longitudinal data, focusing only on early career teachers.

Interestingly, compared to its association with ECTs leaving the school, principal leadership had a weaker association with ECTs leaving the profession. To be sure, previous year's principal leadership still had a significant association, but the magnitude of the association decreased. Moreover, principal leadership in a teacher's first year did not have a significant association with teachers leaving the profession. That is, principal leadership, as part of the organizational context of the school, did not strongly affect ECTs' decisions to leave the profession, but did influence their decisions to leave the current school. Hence, if the matter at hand is losing too many new teachers across the profession, focusing on principal leadership might not be the most effective approach. However, as many studies have documented (e.g., Allensworth et al., 2009; Bryk & Schneider, 2002; Guin, 2004; Ronfeldt et al., 2013), losing a teacher is still a serious issue at the school level. To address this issue, the findings from this paper suggest that principal leadership might be key.

The question becomes how principal leadership influences ECTs' decisions to leave their schools. The second main finding of this study about three aspects of principal leadership provides an answer to this question. Previous studies did not distinguish among different aspects of principal leadership, so that it was hard to specify how principal leadership affects teacher turnover. In contrast, this study conceptualizes principal leadership as having three relevant but distinctive aspects: instructional leadership, principal leadership related to student management, and principal leadership related to creating a supportive culture. After controlling for teachers' and schools' characteristics, principal

leadership related to creating a supportive culture had a significant negative association with teachers' likelihood of leaving the school, but instructional leadership and leadership related to behavior management did not.

Fostering a supportive culture involves different leadership practices than the other two aspects; it is about the relationships among school community members and overall climate in the school rather than the direct relationships between ECTs and school leaders. With a supportive culture in the school, ECTs can raise their concerns to school community members easily and learn through cooperation with other teachers. The findings of this study indicate that this collective support for ECTs might be more important than one principal's direct support for ECTs' everyday tasks, including instructional practice and student behavior management. Instead, as Johnson and colleagues (2012) argued, support from other teachers might be also important for motivating ECTs to stay in the school longer. This finding aligns well with the results from previous studies, in that cooperative working conditions, frequent interactions among teachers, and supportive school cultures were important factors for teacher retention (Allensworth et al., 2009; Borman & Bowling, 2008; Johnson, 2003; Johnson et al., 2012; Simon & Johnson, 2013). The findings also resonate with distributed leadership theory; it focuses not on positionality or formal roles, but on leadership activities that can be conducted by any member of the school community (Spillane, Halverson & Diamond, 2001). Building a supportive culture may include distributed leadership and the findings from this paper suggest that in terms of retaining more ECTs in their schools, individual teachers can play a role as well as principals. Moreover, relational trust underlying such a collaborative school culture has been argued to have a positive effect on student learning (Bryk & Schneider, 2002). Thus, cultivating a supportive culture might be conducive not only to retaining more teachers, but also to improving student learning.

The third main finding is related to the heterogeneous effects of principal leadership. In terms of teachers leaving the school, two variables had significant interaction effects with principal leadership related to creating a supportive culture: teaching at schools located in suburbs and teaching at elementary schools. The influence of principal leadership related to creating a supportive culture was stronger in

suburban schools and weaker in elementary schools. The result regarding suburban schools is quite surprising in that it contrasts with ones from previous studies (i.e., Grissom, 2011; Ladd, 2011). Principal leadership related to creating a supportive culture was expected to have a stronger effect on teacher turnover in urban or rural schools, which are sometimes not capable of providing teachers with enough resources for their teaching. This aspect of principal leadership might enable teachers to share their resources and expertise easily, and such sharing can offset the lack of resources. However, given that this result pertains to ECTs, it may be the case that parents of students in suburban schools are more likely to ask more of teachers and schools (Prater, Bermudez, & Owens, 1997), so ECTs without much experience and expertise might need more help from experienced teachers and/or administrators.

The weaker effect of principal leadership related to a creating a supportive culture in elementary school is also hard to explain. Although elementary schools usually feature self-contained classrooms, teachers at the same grade or adjacent grades might be more influential for ECTs working at elementary schools, compared to their counterparts working at secondary schools where teachers have distinctly different assignments from each other based on their subjects. However, the elementary school environment can naturally facilitate collaboration among teachers, so it is plausible that principals' efforts to create such a supportive environment are not as important. In contrast, collaboration among secondary school teachers might require particular effort or support from principals. In sum, there are some plausible explanations for the heterogeneous effects of principal leadership related to creating a supportive culture on teacher turnover, but further studies with detailed information about each school organization are necessary.

Although the findings of this paper contribute to the literature on teacher turnover, there are some limitations. First, whereas using a broader definition of principal leadership enabled me to disentangle the impact of different aspects of principal leadership on ECT turnover, it is difficult to completely separate the impact of other available resources in schools. For example, while providing additional preparation time for ECTs might be under principals' control, it is also possible that this decision depends on resources for schools, a decision that is under district administrators' control. However, it should be noted

that some school characteristics that might be correlated with the level of available resources, such as students' SES status, school size, and school locations, were included in the models.

Second, principal leadership, the main concern of the paper, was measured only by individual ECTs' reports about their principals. Unfortunately, other relatively more objective measures of principal leadership, such as multiple teachers' judgments about their principal, were not available due to the structure of the BTLS data. Therefore, it is possible that an ECT who does not favor the job or the school in general reported the principal to be a poor leader and left the school. If this is the case, the ECT's decision to leave the school was not directly a function of principal leadership, but instead a result of the ECT's negative overall sentiment about the school. Although it is hard to completely rule out this possibility, ECT commitment level was included as a control variable in order to control for their future career plans and general sentiment toward their jobs. In addition, Robustness Indices were used and two inferences about the association between principal leadership and ECTs leaving the school, in particular, seemed to be quite strong.

Third, the three aspects of principal leadership were only available in the 2007-08 survey. Although time-variant general principal leadership focused on ECTs' current principals, the measures of the three aspects of principal leadership were from the teachers' principals in their first year. That is, it was impossible to examine these three aspects of principal leadership for ECTs in their second through fifth years of teaching. For future research based on longitudinal data about ECTs, using the same measures during each year for five years can address these questions.

Fourth, due to the limited sample size, it was not possible to take into account ECTs' reasons for turnover. Although there were items in the BTLS that asked about ECTs' reasons why they left a school and/or the profession, it was not appropriate to limit the sample based on their reasons. When ECTs who left their school or the profession because of school budget issues or other issues potentially beyond the principals' or ECTs' control were excluded from the model, the sample size was too small. Moreover, it is hard to know whether a reason was purely a budget issue or combination of multiple issues facing an ECT.

In addition, reasons such as these were reported by ECTs, so there is a possibility that the identified reasons themselves are not accurate, as well.

Based on the findings and limitations of the study, I suggest a few directions for future research on the issue of teacher turnover. First, more studies on training principals to reduce teacher turnover are necessary. Although the current study and the previous literature have shown that principal leadership, especially that related to creating a supportive culture, is important for retaining more teachers, how to effectively train principals in this area is unclear. One of the possible directions for research is randomized control trials for evaluating principal leadership training programs in light of reducing teacher turnover. In a similar vein, comparative studies across different districts/states that have different policies and/or support systems for principals could be fruitful. For example, a district that puts particular emphasis on principal leadership related to creating a supportive culture by providing more professional development may be better off in terms of retaining more teachers compared to other districts that did not. Second, some limitations arise due to the unique structure of the BTLS data: 1) three aspects of principal leadership were only available at Year 1; 2) Year 2 data did not have principal leadership data at all; and 3) only Years 3 through 5 shared the same set of survey items about principal leadership. Accordingly, a future research study using more complete data could confirm the findings from the current study, especially for the comparison of the influence of principal leadership across years.

As expected, principal leadership was important for ECT turnover. When an ECT perceived his/her principal as a strong leader, the ECT was less likely to leave the school or the profession across years, even after controlling for various factors that might affect turnover. However, the findings from this paper also suggested that retaining more ECTs is not only the role of principals, but also a mission for the entire school community. Creating a supportive culture in schools may be more difficult than fostering the other two aspects of principal leadership because it involves not only principal-ECT relationships but also relationships among the principal and all teachers in a school. Moreover, it involves not only physical resources or specific activities, but also intangible factors, such as trust and support among teachers and their principal, which can take a great deal of time and social energy to establish. However, creating a

supportive culture can benefit ECTs as well as experienced teachers (Allensworth et al., 2009; Ladd, 2001) and perhaps most importantly, students (Bryk & Schneider, 2002). That is, despite the costs, the effects of a positive culture are clear. Thus, it is important to expand the notion of leadership and emphasize the role of culture in schools. This study is another example about the effects of school culture on one of the most important issues in schools: teacher turnover.

NOTES

NOTES

¹ Since the three aspects of principal leadership were measured at the Year 1, this variable does not have a subscript *j*. For some models, I included principal leadership variables as dummy variables, rather than linear terms. Those dummy variables were not significant in any models, regardless of including them along with linear terms or by themselves.

APPENDICES

Appendix A

Survey Items Used in Analysis

Table 27. Survey items used in analysis

Variable	Survey	Item
General Leadership (year 1)	2007-08 BTLS	Regular supportive communication with your principal, other administrators, or department chair (yes or no) My principal enforces school rules for student conduct and backs me up when I need it. The school administration's behavior toward the staff is supportive and encouraging.
General Leadership (year 3-5)	2009-10 2010-11 2011-12 BTLS	My principal supports me in classroom management issues when I need it. My principal supports me in my interactions with parents when I need it. My principal is approachable My principal listens to my concerns. My principal supports my professional development beyond those activities that are required. My principal has professional respect for teachers. My principal encourages collaboration among teachers.
Principals' instructional leadership	2007-08 BTLS	Did you receive the following kinds of support during your FIRST year of teaching? -Reduced teaching schedule or number of preparations -Common planning time with teachers in your subject -Extra classroom assistance Necessary materials such as textbooks, supplies, and copy machines are available as needed by the staff. Routine duties and paperwork interfere with my job of teaching. In this school, staff members are recognized for a job well done.
Principals' leadership related to managing students' behavior	2007-08 BTLS	The level of student misbehavior in this school (such as noise, horseplay or fighting in the halls, cafeteria, or student lounge) interferes with my teaching. My principal enforces school rules for student conduct and backs me up when I need it. Rules for student behavior are consistently enforced by teachers in this school, even for students who are not in their classes. The amount of student tardiness and class cutting in this school interferes with my teaching.

Table 27 (cont'd)

Variable	Survey	Item
Principals' leadership related to fostering supportive culture	2007-08 BTLS	Did you receive the following kinds of support during your FIRST year of teaching? -Regular supportive communication with your principal, other administrators, or department chair The school administration's behavior toward the staff is supportive and encouraging Most of my colleagues share my beliefs and values about what the central mission of the school should be The principal knows what kind of school he or she wants and has communicated it to the staff There is a great deal of cooperative effort among the staff members I like the way things are run at this school
Induction	2007-08 BTLS	In your first year of teaching, did you participate in a teacher induction program?
PD	2007-08 BTLS	In the past 12 months, how many hours did you spend on these activities? - Content specific, Computers, Reading, discipline, special education, teaching students with limited-English proficiency
Work hours	2007-08 BTLS	Including hours spent during the school day, before and after school, and on the weekends, how many hours do you spend on ALL teaching and other school-related activities during a typical full week at this school?
Autonomy	2007-08 BTLS	How much actual control do you have in your classroom at this school over the following areas of your planning and teaching? - Selecting textbooks and other instructional materials - Selecting content, topics, and skills to be taught - Selecting teaching techniques - Evaluating and grading students - Disciplining students - Determining the amount of homework to be assigned
Salary	All waves of BTLS	During the current school year, what is your academic year base teaching salary?
Parents' involvement	2007-08 SASS principal	Last school year (2006-07), what percentage of students had at least one parent or guardian participating in the following events? - Open house or back-to-school night - All regularly scheduled schoolwide parent-teacher conferences - One or more special subject-area events -Volunteer in the school on a regular basis

Table 27 (cont'd)

Variable	Survey	Item
School safety	2007-08 SASS principal	To the best of your knowledge, how often do the following types of problems occur at this school? - Physical conflicts among students - Robbery or theft - Vandalism - Student use of alcohol - Student use of illegal drugs - Student possession of weapons - Physical abuse of teachers - Student racial tensions - Student Bullying - Student verbal abuse of teachers - Widespread disorder in classrooms - Student acts of disrespect for teachers - Gang activity
AYP status	2007-08 SASS principal	At the end of the last school year (2006-07), did this school make Adequate Yearly Progress (AYP)?

Appendix B

Correlation Between Weights and Principal Leadership and Control Variables

Table 28. Correlation Between Weights and Principal Leadership and Control Variables

	Leaving the school weights	Leaving the profession weights
General leadership	0.0231* (0.0397)	0.0257* (0.022)
Instructional leadership	0.0102 (0.3109)	0.0122 (0.2256)
Leadership related to student management	0.0979*** (≤0.0001)	0.0919*** (≤0.0001)
Leadership related to supportive culture	0.0152 (0.1291)	0.008 (0.4271)
Preparation	0.0159 (0.1178)	0.0203* (0.0468)
Alternative certificate	-0.0047 (0.6357)	-0.0043 (0.6647)
Induction	0.1112*** (≤0.0001)	0.1072*** (≤0.0001)
Work hours	0.0475*** (≤0.0001)	0.0407*** (0.0001)
Autonomy	-0.0893*** (≤0.0001)	-0.0818*** (≤0.0001)
Male	-0.0641*** (≤0.0001)	-0.072*** (≤0.0001)
White	0.0191 (0.0566)	0.0164 (0.1009)
Union membership	0.047*** (≤0.0001)	0.0429*** (≤0.0001)
Advanced degree	-0.0051 (0.6103)	-0.01 (0.3202)
Salary(log)	0.1396*** (≤0.0001)	0.1392*** (≤0.0001)
HQT	0.0434** (0.0002)	0.0459*** (0.0001)
PD	0.0867*** (≤0.0001)	0.0894*** (≤0.0001)
Demanding subject	-0.0464*** (≤0.0001)	-0.0476*** (≤0.0001)
Commitment	0.0534*** (≤0.0001)	0.0397*** (0.0001)
School size	0.0231* (0.0214)	0.0266** (0.008)
% of racially minority students	0.0996*** (≤0.0001)	0.1147*** (≤0.0001)
School safety	0.1022*** (≤0.0001)	0.1074*** (≤0.0001)
Parents' involvement	0.1288*** (≤0.0001)	0.1249*** (≤0.0001)

Table 28 (cont'd)

	Leaving the school weights	Leaving the profession weights
AYP status	0.0553*** (≤ 0.0001)	0.0432*** (≤ 0.0001)
% of FRL	0.0214 (0.0668)	0.0234* (0.0443)
Suburban	0.0744*** (≤ 0.0001)	0.0775*** (≤ 0.0001)
Charter school	-0.031** (0.0064)	-0.0351** (0.0026)
Elementary school	0.261*** (≤ 0.0001)	0.268*** (≤ 0.0001)
High school	-0.1946*** (≤ 0.0001)	-0.1999*** (≤ 0.0001)

Appendix C

Analysis Including Interaction Terms Between Weights and Principal Leadership Variables

Table 29. The Influence of Principal Leadership on Leaving the School (Including interaction terms)

	Model 1	Model 2	Model 3	Model 4
General leadership (1year lagged)	0.701*** (0.0614)			
General leadership *weight	1.000 (0.000118)			
Instructional Leadership		0.831 (0.12)		
Instructional leadership *weight		1.000 (0.000207)		
Leadership related to student management			0.908 (0.0765)	
Leadership related to student management *weight			1.000* (0.000106)	
Leadership related to supportive culture				0.687** (0.0728)
Leadership related to supportive culture *weight				1.000* (0.000119)
Preparation	0.856 (0.0736)	0.861 (0.0741)	0.860 (0.0739)	0.860 (0.0738)
Alternative certificate	0.929 (0.102)	0.930 (0.101)	0.933 (0.102)	0.935 (0.103)
Induction	0.938 (0.0989)	0.939 (0.0994)	0.936 (0.0989)	0.950 (0.101)
Work hours	0.993 (0.00421)	0.993 (0.00426)	0.993 (0.00424)	0.993 (0.00424)
Autonomy	0.878 (0.0834)	0.842 (0.0800)	0.840 (0.0799)	0.881 (0.0850)
Male	0.910 (0.0895)	0.895 (0.0879)	0.882 (0.0866)	0.907 (0.0889)
White	0.753 (0.121)	0.762 (0.119)	0.766 (0.123)	0.756 (0.119)
Union Membership	0.922 (0.0880)	0.943 (0.0898)	0.944 (0.0897)	0.925 (0.0885)

Table 29 (cont'd)

	Model 1	Model 2	Model 3	Model 4
Advanced degree	1.149 (0.133)	1.145 (0.133)	1.167 (0.135)	1.159 (0.134)
Salary(log)	0.484*** (0.0856)	0.489*** (0.0864)	0.496*** (0.0876)	0.482*** (0.0850)
HQT	0.881 (0.0963)	0.882 (0.0965)	0.888 (0.0972)	0.879 (0.0962)
PD	1.043 (0.0228)	1.040 (0.0228)	1.041 (0.0229)	1.046* (0.0229)
Demanding subject	0.986 (0.0969)	0.993 (0.0974)	0.986 (0.0969)	0.990 (0.0969)
Commitment	0.589*** (0.0499)	0.562*** (0.0481)	0.561*** (0.0481)	0.601*** (0.0523)
School size	1.000 (0.0000937)	1.000 (0.0000933)	1.000 (0.0000938)	1.000 (0.0000936)
% of racially minority students	1.005** (0.00184)	1.006** (0.00183)	1.005** (0.00184)	1.005** (0.00185)
School safety	1.001 (0.116)	0.985 (0.114)	1.005 (0.117)	1.017 (0.119)
Parents' involvement	0.961 (0.0695)	0.953 (0.0690)	0.961 (0.0694)	0.954 (0.0693)
AYP status	0.804* (0.0803)	0.798* (0.0794)	0.799* (0.0799)	0.816* (0.0815)
% of FRL	0.998 (0.00252)	0.998 (0.00251)	0.998 (0.00251)	0.998 (0.00252)
Suburb	0.948 (0.112)	0.962 (0.113)	0.960 (0.114)	0.981 (0.116)
Charter school	0.823 (0.182)	0.830 (0.184)	0.819 (0.182)	0.813 (0.181)
Elementary school	1.134 (0.155)	1.138 (0.155)	1.162 (0.159)	1.157 (0.159)
High school	0.913 (0.113)	0.878 (0.109)	0.875 (0.108)	0.890 (0.110)
N	3760	3760	3760	3760

Note. Weights are the leaving the school weights (not transformed). Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 30. The Influence of Principal Leadership on Leaving the Profession(Including interaction terms)

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.792*** (0.0891)			
General leadership *weight	1.000 (0.000186)			
Instructional Leadership		0.738 (0.153)		
Instructional leadership *weight		0.999 (0.000337)		
Leadership related to student management			0.823 (0.0919)	
Leadership related to student management *weight			1.000* (0.000167)	
Leadership related to supportive culture				0.883 (0.128)
Leadership related to supportive culture *weight				1.000* (0.000189)
Preparation	0.635*** (0.0730)	0.644*** (0.0740)	0.643*** (0.0738)	0.637*** (0.0731)
Alternative Certificate Induction	0.844 (0.130)	0.845 (0.129)	0.850 (0.130)	0.852 (0.131)
	0.853 (0.121)	0.849 (0.120)	0.843 (0.119)	0.844 (0.119)
Work hours	0.989* (0.00567)	0.988* (0.00580)	0.988* (0.00580)	0.989 (0.00569)
Autonomy	0.936 (0.122)	0.904 (0.117)	0.908 (0.117)	0.906 (0.118)
Male	0.998 (0.140)	0.973 (0.136)	0.951 (0.133)	0.968 (0.135)
White	0.749 (0.162)	0.750 (0.160)	0.753 (0.163)	0.761 (0.163)
Union Membership	0.908 (0.121)	0.942 (0.125)	0.940 (0.125)	0.941 (0.125)
Advanced Degree	1.150 (0.189)	1.129 (0.184)	1.153 (0.189)	1.154 (0.189)
Salary(log)	0.404*** (0.0892)	0.411*** (0.0897)	0.411*** (0.0901)	0.415*** (0.0905)

Table 30 (cont'd)

	Model 1	Model 2	Model 3	Model 4
HQT	0.677** (0.0991)	0.680** (0.0995)	0.684** (0.100)	0.681** (0.0999)
PD	1.053 (0.0310)	1.049 (0.0306)	1.053 (0.0307)	1.051 (0.0306)
Demanding Subject Commitment	1.036 (0.138) 0.614*** (0.0701)	1.039 (0.138) 0.600*** (0.0700)	1.028 (0.137) 0.602*** (0.0689)	1.033 (0.137) 0.597*** (0.0711)
School size	1.000 (0.000135)	1.000 (0.000134)	1.000 (0.000136)	1.000 (0.000135)
% of racially minority students	1.005 (0.00248)	1.005* (0.00244)	1.005 (0.00247)	1.005* (0.00248)
School safety	1.217 (0.186)	1.158 (0.180)	1.191 (0.186)	1.168 (0.181)
Parents' Involvement	1.008 (0.101)	1.007 (0.1000)	1.015 (0.101)	1.005 (0.100)
AYP status	0.930 (0.130)	0.942 (0.131)	0.944 (0.131)	0.938 (0.130)
% of FRL	0.997 (0.00330)	0.996 (0.00330)	0.996 (0.00329)	0.996 (0.00331)
Suburban	0.922 (0.156)	0.939 (0.157)	0.921 (0.154)	0.937 (0.157)
Charter school	1.123 (0.301)	1.188 (0.316)	1.172 (0.309)	1.177 (0.313)
Elementary School	0.944 (0.184)	0.946 (0.183)	0.963 (0.185)	0.956 (0.185)
High school	1.258 (0.214)	1.195 (0.204)	1.189 (0.204)	1.210 (0.207)
N	4320	4330	4320	4330

Note. Weights are the leaving the profession weights (not transformed). Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The coefficients are reported as odds ratios. The standard errors were clustered at individual the teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix D

Analysis Including Interaction Terms Between Race/School Size/Ratio of Racially Minority Students and Principal Leadership Variables Without Weights

Table 31. The Influence of Principal Leadership on Leaving the School Without Weights

Panel A. Principal leadership*Race

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.687*** (0.0596)			
Instructional leadership		0.809 (0.116)		
Leadership related to student management			0.877 (0.0716)	
Leadership related to supportive culture				0.671*** (0.0705)
White	0.752 (0.121)	0.78 (0.124)	0.767 (0.119)	0.771 (0.122)
White* Leadership	0.743 (0.172)	1.643 (0.727)	0.822 (0.19)	1.005 (0.307)
N	3760	3760	3760	3760

Panel B. Principal leadership*School size

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.687*** (0.0596)			
Instructional leadership		0.807 (0.115)		
Leadership related to student management			0.877 (0.0716)	
Leadership related to supportive culture				0.665*** (0.0693)
School Size	1 (0.0001)	1 (0.0001)	1 (0.0001)	1 (0.0001)
School size*Leadership	1 (0.0001)	1 (0.0002)	1 (0.0001)	1 (0.0001)
N	3760	3760	3760	3760

Panel C. Principal leadership*% of racially minority students

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.683*** (0.059)			
Instructional leadership		0.806 (0.115)		
Leadership related to student management			0.876 (0.072)	
Leadership related to supportive culture				0.670*** (0.0704)
% of racially minority students	1.005** (0.0019)	1.005** (0.00184)	1.005** (0.00184)	1.005** (0.00185)

Table 31 (cont'd)

	Model 1	Model 2	Model 3	Model 4
% of racially minority students *Leadership	1.002 (0.0021)	0.995 (0.00368)	1.001 (0.00198)	1.001 (0.00260)
N	3760	3760	3760	3760

Table 32. The Influence of Principal Leadership on Leaving the Profession Without Weights

<i>Panel A. Principal leadership*Race</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.675*** (0.0727)			
Instructional leadership		0.702 (0.146)		
Leadership related to student management			0.790* (0.0879)	
Leadership related to supportive culture				0.810 (0.124)
White	0.675 (0.145)	0.750 (0.163)	0.736 (0.157)	0.773 (0.167)
White* Leadership	0.605 (0.178)	1.036 (0.592)	0.846 (0.249)	1.189 (0.511)
N	4320	4330	4320	4330
<i>Panel B. Principal leadership*School size</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.671*** (0.0743)			
Instructional leadership		0.722 (0.153)		
Leadership related to student management			0.805 (0.091)	
Leadership related to supportive culture				0.810 (0.126)
School Size	1 (0.000138)	1 (0.00014)	1 (0.000136)	1 (0.000135)
School size*Leadership	1 (0.00015)	1 (0.000314)	1 (0.000179)	1 (0.000247)
N	4320	4330	4320	4330
<i>Panel C. Principal leadership*% of racially minority students</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.672*** (0.0743)			
Instructional leadership		0.705 (0.146)		
Leadership related to student management			0.799* (0.089)	
Leadership related to supportive culture				0.812 (0.124)
% of racially minority students	1.004 (0.00254)	1.005 (0.00247)	1.004 (0.00251)	1.004 (0.0025)
% of racially minority students *Leadership	1 (0.00244)	0.998 (0.00488)	0.999 (0.00263)	0.998 (0.00363)
N	4320	4330	4320	4330

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were

Table 32 (cont'd)
 included). The variables for calculating interaction terms were centered by their grand means. The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix E

Results Using Replicate Weights Instead of Teacher Clustered Errors

Table 33. The Influence of Principal Leadership on ECT Leaving the School and Leaving the Profession (Using replicate weights)

	Leaving the school				Leaving the profession			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.7 (0.142)				0.731 (0.192)			
Instructional leadership		0.680 (0.184)				0.857 (0.279)		
Leadership related to student management			0.885 (0.281)				0.871 (0.208)	
Leadership related to supportive culture				0.813 (0.309)				0.784 (0.42)
Preparation	0.853 (0.179)	0.852 (0.181)	0.854 (0.175)	0.857 (0.170)	0.676 (0.172)	0.675 (0.170)	0.680 (0.170)	0.682 (0.177)
Alternative certificate	0.950 (0.300)	0.951 (0.283)	0.952 (0.275)	0.948 (0.287)	0.825 (0.432)	0.829 (0.434)	0.828 (0.422)	0.825 (0.431)
Induction	0.919 (0.267)	0.930 (0.269)	0.914 (0.261)	0.919 (0.261)	0.823 (0.257)	0.816 (0.253)	0.814 (0.247)	0.820 (0.247)
Work hours	0.990 (0.0108)	0.990 (0.0111)	0.990 (0.0108)	0.990 (0.0113)	0.989 (0.0115)	0.989 (0.0116)	0.989 (0.0118)	0.988 (0.0124)
Autonomy	0.869 (0.209)	0.872 (0.209)	0.834 (0.208)	0.834 (0.203)	0.958 (0.268)	0.931 (0.266)	0.927 (0.259)	0.928 (0.264)
Male	0.878 (0.205)	0.875 (0.205)	0.850 (0.220)	0.862 (0.209)	0.980 (0.343)	0.958 (0.337)	0.945 (0.327)	0.959 (0.336)
White	0.731 (0.310)	0.734 (0.313)	0.746 (0.325)	0.743 (0.320)	0.747 (0.347)	0.757 (0.352)	0.753 (0.352)	0.752 (0.345)
Union membership	0.964 (0.238)	0.962 (0.233)	0.983 (0.241)	0.984 (0.241)	0.973 (0.298)	0.998 (0.300)	1.001 (0.308)	1.002 (0.300)
Advanced degree	1.027 (0.372)	1.037 (0.390)	1.044 (0.388)	1.026 (0.378)	1.136 (0.430)	1.144 (0.419)	1.143 (0.430)	1.125 (0.424)
Salary(log)	0.532 (0.281)	0.522 (0.279)	0.542 (0.298)	0.535 (0.289)	0.356 (0.224)	0.361 (0.225)	0.361 (0.228)	0.359 (0.228)
HQT	0.937 (0.203)	0.926 (0.210)	0.939 (0.215)	0.934 (0.212)	0.775 (0.250)	0.774 (0.253)	0.776 (0.255)	0.775 (0.253)
PD	1.038 (0.0747)	1.042 (0.0764)	1.037 (0.0730)	1.036 (0.0741)	1.053 (0.0782)	1.051 (0.0800)	1.052 (0.0782)	1.050 (0.0774)
Demanding subject	1.031 (0.285)	1.034 (0.287)	1.039 (0.277)	1.045 (0.276)	1.109 (0.382)	1.107 (0.377)	1.104 (0.374)	1.116 (0.374)
Commitment	0.603 (0.155)	0.613* (0.147)	0.576 (0.162)	0.573* (0.142)	0.590* (0.151)	0.573* (0.146)	0.573* (0.138)	0.574* (0.140)

Table 33 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
School size	1.000 (0.0002)	1.000 (0.0002)	1.000 (0.0002)	1.000 (0.0002)	1.000 (0.0004)	1.000 (0.0004)	1.000 (0.0004)	1.000 (0.0004)
% of racially minority students	1.003 (0.0045)	1.003 (0.0044)	1.004 (0.0044)	1.004 (0.0045)	1.003 (0.0058)	1.003 (0.0058)	1.003 (0.0059)	1.003 (0.0056)
School safety	0.935 (0.288)	0.946 (0.285)	0.940 (0.307)	0.919 (0.276)	1.118 (0.701)	1.079 (0.646)	1.093 (0.647)	1.072 (0.648)
Parent involvement	0.924 (0.160)	0.920 (0.164)	0.921 (0.160)	0.913 (0.158)	0.914 (0.215)	0.910 (0.208)	0.918 (0.210)	0.911 (0.213)
AYP status	0.762 (0.210)	0.776 (0.219)	0.757 (0.210)	0.758 (0.211)	0.861 (0.271)	0.865 (0.271)	0.865 (0.270)	0.866 (0.287)
% of FRL	1.000 (0.0094)	1.000 (0.0095)	1.000 (0.009)	1.000 (0.0095)	0.997 (0.0071)	0.997 (0.0073)	0.996 (0.0072)	0.997 (0.0071)
Suburban	0.937 (0.326)	0.969 (0.342)	0.940 (0.342)	0.944 (0.322)	0.855 (0.307)	0.860 (0.313)	0.852 (0.303)	0.858 (0.308)
Charter school	0.706 (0.365)	0.698 (0.368)	0.704 (0.394)	0.714 (0.364)	0.980 (0.498)	1.024 (0.522)	1.020 (0.525)	1.034 (0.522)
Elementary school	1.139 (0.427)	1.144 (0.431)	1.153 (0.409)	1.139 (0.427)	0.941 (0.467)	0.940 (0.462)	0.949 (0.449)	0.939 (0.435)
High school	0.881 (0.261)	0.864 (0.256)	0.843 (0.247)	0.846 (0.254)	1.132 (0.522)	1.100 (0.498)	1.086 (0.485)	1.089 (0.485)
N	5340	5340	5340	5340	5800	5810	5800	5810

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The coefficients are reported as odds ratios and the estimates were weighted by log of weights. BTLS 5th year longitudinal replicate weights (i.e., 1-88) were applied for obtaining standard errors. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * p<0.05, ** p<0.01, *** p<0.001

Appendix F

The Results Using Untransformed Weights

Table 34. The Influence of Principal Leadership on Leaving the School (Using untransformed weights)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.656*** (0.0725)	0.748* (0.0948)				0.662** (0.0835)	0.769 (0.107)	
General leadership (2 year lagged)		0.924 (0.123)					0.868 (0.122)	
Instructional leadership			0.904 (0.155)					0.799 (0.205)
Leadership related to student management				0.673*** (0.0729)				
Leadership related to supportive culture					0.656*** (0.106)			
Preparation						0.84 (0.113)	0.781 (0.12)	0.843 (0.113)
Alternative certificate						1.193 (0.226)	1.346 (0.274)	1.213 (0.226)
Induction						0.848 (0.145)	0.879 (0.157)	0.856 (0.151)
Work hours						0.986 (0.00812)	0.991 (0.00859)	0.986 (0.00828)
Autonomy						0.878 (0.14)	0.832 (0.148)	0.837 (0.133)
Male						0.986 (0.165)	1.149 (0.214)	0.951 (0.161)
White						0.716 (0.182)	0.659 (0.192)	0.727 (0.187)
Union membership						0.991 (0.154)	1.032 (0.181)	1.028 (0.16)
Advanced Degree						0.854 (0.184)	0.926 (0.207)	0.852 (0.184)
Salary(log)						0.847 (0.267)	0.743 (0.239)	0.88 (0.282)

Table 34 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
HQT						1.066 (0.181)	1.066 (0.197)	1.065 (0.181)
PD						1.013 (0.0449)	1.012 (0.0443)	1.014 (0.0441)
Demanding subject						1.075 (0.184)	1.176 (0.221)	1.097 (0.183)
Commitment						0.805 (0.127)	0.97 (0.162)	0.758 (0.122)
N	4400	3750	4410	4400	4410	3970	3370	3970
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
General leadership (1 yr lagged)			0.718* (0.0993)	0.867 (0.141)				
General leadership (2 yr lagged)				0.86 (0.137)				
Instructional leadership					0.79 (0.205)			1.277 (0.376)
Leadership related to student management	0.702** (0.0949)					0.679* (0.104)		0.754 (0.131)
Leadership related to supportive culture		0.568*** (0.0948)					0.614* (0.119)	0.649 (0.148)
Preparation	0.846 (0.115)	0.841 (0.114)	0.770 (0.109)	0.693* (0.112)	0.775 (0.109)	0.785 (0.110)	0.768 (0.108)	0.769 (0.11)
Alternative certificate	1.196 (0.222)	1.218 (0.228)	1.095 (0.220)	1.247 (0.267)	1.104 (0.218)	1.125 (0.220)	1.130 (0.221)	1.141 (0.211)
Induction	0.843 (0.14)	0.886 (0.155)	0.906 (0.158)	0.915 (0.163)	0.910 (0.161)	0.888 (0.150)	0.930 (0.168)	0.908 (0.159)
Work hours	0.986 (0.00798)	0.987 (0.00823)	0.983* (0.00822)	0.988 (0.00870)	0.983* (0.00839)	0.983* (0.00836)	0.984 (0.00831)	0.984 (0.009)
Autonomy	0.859 (0.137)	0.898 (0.143)	0.895 (0.167)	0.831 (0.176)	0.866 (0.161)	0.890 (0.164)	0.916 (0.170)	0.919 (0.172)
Male	0.883 (0.152)	0.987 (0.168)	0.800 (0.145)	0.939 (0.188)	0.776 (0.142)	0.730 (0.134)	0.810 (0.149)	0.769 (0.141)
White	0.759 (0.196)	0.728 (0.183)	0.778 (0.212)	0.701 (0.227)	0.798 (0.219)	0.800 (0.229)	0.767 (0.206)	0.778 (0.216)

Table 34 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Union membership	1.014 (0.156)	0.975 (0.152)	0.989 (0.171)	1.025 (0.201)	1.020 (0.177)	0.996 (0.170)	0.969 (0.168)	0.958 (0.163)
Advanced Degree	0.868 (0.182)	0.817 (0.181)	0.890 (0.200)	0.954 (0.224)	0.892 (0.202)	0.888 (0.198)	0.870 (0.199)	0.877 (0.199)
Salary(log)	0.872 (0.273)	0.807 (0.252)	0.703 (0.246)	0.631 (0.228)	0.704 (0.249)	0.752 (0.267)	0.658 (0.231)	0.705 (0.252)
HQT	1.074 (0.182)	1.055 (0.18)	0.960 (0.176)	0.962 (0.195)	0.960 (0.176)	0.945 (0.173)	0.945 (0.175)	0.936 (0.173)
PD	1.023 (0.0429)	1.022 (0.0455)	1.015 (0.0447)	1.016 (0.0445)	1.016 (0.0444)	1.020 (0.0437)	1.023 (0.0457)	1.025 (0.045)
Demanding subject	1.058 (0.178)	1.032 (0.174)	1.053 (0.191)	1.161 (0.239)	1.070 (0.191)	1.053 (0.188)	1.016 (0.181)	1.012 (0.181)
Commitment	0.816 (0.131)	0.849 (0.138)	0.715* (0.118)	0.847 (0.150)	0.685* (0.114)	0.738 (0.122)	0.747 (0.128)	0.789 (0.131)
School size			1.000 (0.000154)	1.000 (0.000170)	1.000 (0.000159)	1.000 (0.000153)	1.000 (0.000151)	1.000 (0.00015)
% of racially minority students			1.002 (0.00313)	1.002 (0.00343)	1.003 (0.00307)	1.002 (0.00307)	1.001 (0.00315)	1.001 (0.003)
School safety			0.826 (0.171)	0.765 (0.178)	0.824 (0.171)	0.899 (0.190)	0.859 (0.180)	0.910 (0.190)
Parents' involvement			0.842 (0.113)	0.809 (0.122)	0.829 (0.112)	0.857 (0.115)	0.846 (0.113)	0.863 (0.115)
AYP status			0.762 (0.132)	0.741 (0.146)	0.768 (0.133)	0.777 (0.134)	0.793 (0.138)	0.793 (0.137)
% of FRL			1.008 (0.00467)	1.007 (0.00517)	1.008 (0.00468)	1.007 (0.00457)	1.008 (0.00470)	1.008 (0.00467)
Suburban			1.111 (0.223)	1.050 (0.234)	1.107 (0.223)	1.137 (0.225)	1.195 (0.242)	1.2001 (0.244)
Charter school			0.600 (0.195)	0.446* (0.172)	0.614 (0.200)	0.568 (0.186)	0.589 (0.195)	0.559 (0.185)
Elementary school			1.090 (0.275)	1.182 (0.335)	1.093 (0.272)	1.128 (0.273)	1.076 (0.268)	1.094 (0.266)
High school			0.728 (0.160)	0.774 (0.192)	0.697 (0.156)	0.706 (0.152)	0.739 (0.163)	0.736 (0.16)

Table 34 (cont'd)

N	3960	3970	3510	2980	3510	3500	3510	3500
---	------	------	------	------	------	------	------	------

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). Estimates are adjusted by leaving the school weights (i.e., not transformed weights). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 35. The Heterogeneous Effects of Principal Leadership on Leaving the School (Using untransformed weights)

<i>Panel A. Principal leadership*time indicators</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.814 (0.132)			
Instructional leadership		0.710 (0.216)		
Leadership related to student management			0.662* (0.112)	
Leadership related to supportive culture				0.555* (0.137)
Year 2	1080.8 (3872.5)	1538.3* (5757.5)	834.9 (3118.2)	5422.1* (20173.2)
Year 3	621.1 (2244.8)	785.1 (2961.5)	423.8 (1595.2)	2791.8* (10464.0)
Year 4	780.6 (2839.0)	739.7 (2803.3)	424.5 (1605.6)	2706.4* (10202.4)
Year 5	673.3 (2454.0)	764.9 (2904.8)	428.9 (1623.0)	2735.6* (10337.0)
Year 2* Leadership	0.498 (0.187)	1.261 (0.762)	1.079 (0.337)	1.368 (0.757)
Year 3* Leadership	0.646 (0.305)	2.766 (2.104)	1.268 (0.459)	2.206 (1.317)
Year 4* Leadership	0.566 (0.214)	0.805 (0.623)	1.062 (0.387)	1.075 (0.693)
N	3510	3510	3500	3510
<i>Panel B. Principal leadership*Suburban</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.715* (0.0993)			
Instructional leadership		0.790 (0.205)		
Leadership related to student management			0.669** (0.102)	
Leadership related to supportive culture				0.617* (0.121)
Suburban	1.163 (0.246)	1.106 (0.224)	1.127 (0.219)	1.190 (0.239)
Suburban*Leadership	1.212 (0.330)	0.970 (0.475)	1.216 (0.334)	0.563 (0.220)
N	3510	3510	3500	3510

Table 35 (cont'd)

*Panel C. Principal leadership*Elementary school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.668*** (0.0905)			
Instructional leadership		0.715 (0.194)		
Leadership related to student management			0.650** (0.0971)	
Leadership related to supportive culture				0.550*** (0.108)
Elementary school	1.171 (0.292)	1.102 (0.273)	1.083 (0.268)	1.080 (0.268)
Elementary school*Leadership	1.407 (0.318)	1.448 (0.645)	1.295 (0.354)	1.599 (0.488)
N	3510	3510	3500	3510

*Panel D. Principal leadership*High school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.702** (0.0921)			
Instructional leadership		0.761 (0.180)		
Leadership related to student management			0.682** (0.0989)	
Leadership related to supportive culture				0.550** (0.101)
High school	0.713 (0.161)	0.689 (0.156)	0.708 (0.154)	0.72 (0.161)
High school*Leadership	0.912 (0.190)	0.867 (0.348)	1.028 (0.244)	0.664 (0.189)
N	3510	3510	3500	3510

*Panel E. Principal leadership*Charter school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.717* (0.0987)			
Instructional leadership		0.788 (0.204)		
Leadership related to student management			0.680* (0.105)	
Leadership related to supportive culture				0.613** (0.119)
Charter school	0.487 (0.183)	0.563 (0.202)	0.551 (0.187)	0.614 (0.208)
Charter school* Leadership	0.627 (0.233)	0.551 (0.366)	0.834 (0.322)	1.241 (0.599)
N	3510	3510	3500	3510

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible

Table 35 (cont'd)

students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The variables for calculating interaction terms were centered by their grand means. Estimates are adjusted by leaving the school weights (i.e., not transformed weights). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 36. The Influence of Principal Leadership on Leaving the Profession (Using untransformed weights)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.734* (0.101)	0.743 (0.127)				0.756 (0.119)	0.753 (0.142)	
General leadership (2 year lagged)		0.941 (0.183)						0.547 (0.175)
Instructional leadership			0.734 (0.157)					
Leadership related to student management				0.746* (0.107)				
Leadership related to supportive culture					0.754 (0.129)			
Preparation						0.725 (0.124)	0.733 (0.127)	0.758 (0.131)
Alternative certificate						0.775 (0.171)	0.773 (0.185)	0.793 (0.172)
Induction						0.793 (0.167)	0.802 (0.182)	0.804 (0.168)
Work hours						0.993 (0.00858)	0.996 (0.00963)	0.99 (0.00862)
Autonomy						1.151 (0.219)	1.116 (0.217)	1.165 (0.221)
Male						1.188 (0.259)	1.2 (0.279)	1.164 (0.252)
White						0.873 (0.278)	0.932 (0.348)	0.869 (0.275)
Union membership						1.114 (0.215)	1.060 (0.225)	1.123 (0.215)
Advanced Degree						1.09 (0.289)	1.051 (0.289)	1.044 (0.284)
Salary(log)						0.329*** (0.107)	0.302*** (0.106)	0.322*** (0.105)
HQT						0.864 (0.201)	1.069 (0.28)	0.862 (0.201)
PD						1.024 (0.0542)	1.032 (0.0577)	1.029 (0.0543)

Table 36 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Demanding subject						1.535*	1.713*	1.536*
Commitment						(0.335)	(0.404)	(0.326)
						0.661*	0.769	0.686
						(0.114)	(0.146)	(0.132)
N	5290	4630	5320	5300	5320	4770	4170	4780
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
General leadership (1 year lagged)			0.789 (0.149)	0.762 (0.178)				
General leadership (2 year lagged)				0.995 (0.258)				
Instructional leadership					0.765 (0.255)			0.701 (0.264)
Leadership related to student management	0.861 (0.146)					0.945 (0.184)		0.970 (0.218)
Leadership related to supportive culture		0.781 (0.183)					0.976 (0.241)	1.147 (0.336)
Preparation	0.729 (0.125)	0.721 (0.123)	0.661* (0.125)	0.670* (0.130)	0.671* (0.127)	0.660* (0.124)	0.656* (0.124)	0.678* (0.128)
Alternative certificate	0.787 (0.172)	0.785 (0.173)	0.689 (0.181)	0.650 (0.187)	0.689 (0.180)	0.692 (0.180)	0.689 (0.180)	0.688 (0.179)
Induction	0.792 (0.168)	0.802 (0.171)	0.914 (0.218)	0.887 (0.229)	0.924 (0.221)	0.917 (0.221)	0.918 (0.221)	0.923 (0.221)
Work hours	0.993 (0.00863)	0.993 (0.00868)	0.988 (0.00891)	0.990 (0.0102)	0.987 (0.00915)	0.989 (0.00900)	0.989 (0.00899)	0.987 (0.00906)
Autonomy	1.113 (0.212)	1.139 (0.222)	1.044 (0.240)	0.955 (0.219)	1.026 (0.230)	0.997 (0.228)	0.994 (0.226)	1.016 (0.230)
Male	1.133 (0.237)	1.183 (0.258)	1.054 (0.256)	1.026 (0.267)	1.036 (0.250)	1.025 (0.243)	1.034 (0.250)	1.022 (0.240)
White	0.885 (0.279)	0.875 (0.277)	0.730 (0.241)	0.751 (0.287)	0.744 (0.244)	0.747 (0.247)	0.748 (0.248)	0.751 (0.250)
Union membership	1.142 (0.22)	1.112 (0.211)	1.056 (0.229)	0.957 (0.232)	1.076 (0.233)	1.082 (0.234)	1.079 (0.233)	1.091 (0.236)
Advanced Degree	1.108 (0.295)	1.08 (0.294)	1.034 (0.316)	1.012 (0.324)	1.032 (0.317)	1.057 (0.322)	1.059 (0.321)	1.038 (0.317)

Table 36 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Salary(log)	0.33*** (0.107)	0.329*** (0.107)	0.322*** (0.107)	0.275*** (0.0964)	0.318*** (0.105)	0.322*** (0.107)	0.322*** (0.106)	0.319*** (0.105)
HQT	0.86 (0.199)	0.863 (0.202)	0.887 (0.239)	1.129 (0.341)	0.889 (0.240)	0.884 (0.236)	0.889 (0.238)	0.888 (0.238)
PD	1.026 (0.0542)	1.028 (0.0551)	1.035 (0.0565)	1.045 (0.0548)	1.037 (0.0563)	1.035 (0.0562)	1.035 (0.0563)	1.035 (0.561)
Demanding subject	1.521 (0.331)	1.51* (0.313)	1.082 (0.236)	1.173 (0.284)	1.090 (0.234)	1.085 (0.234)	1.087 (0.230)	1.102 (0.234)
Commitment	0.646* (0.12)	0.66* (0.134)	0.552** (0.105)	0.626* (0.131)	0.546** (0.113)	0.531** (0.106)	0.525** (0.111)	0.540** (0.117)
School size			1.000 (0.000221)	1.000 (0.000254)	1.000 (0.000223)	1.000 (0.000222)	1.000 (0.000225)	1.000 (0.000218)
% of racially minority students			1.002 (0.00442)	1.003 (0.00470)	1.002 (0.00430)	1.002 (0.00435)	1.002 (0.00431)	1.002 (0.004)
School safety			0.694 (0.168)	0.758 (0.194)	0.690 (0.167)	0.696 (0.169)	0.689 (0.166)	0.690 (0.168)
Parents' involvement			0.768 (0.138)	0.772 (0.160)	0.762 (0.135)	0.765 (0.136)	0.762 (0.135)	0.761 (0.135)
AYP status			0.730 (0.167)	0.748 (0.189)	0.737 (0.168)	0.728 (0.165)	0.726 (0.166)	0.732 (0.169)
% of FRL			0.996 (0.00491)	0.998 (0.00539)	0.996 (0.00494)	0.996 (0.00493)	0.996 (0.00497)	0.996 (0.00494)
Suburban			0.787 (0.207)	0.761 (0.216)	0.787 (0.207)	0.789 (0.206)	0.786 (0.206)	0.779 (0.206)
Charter school			1.110 (0.463)	0.733 (0.381)	1.148 (0.474)	1.151 (0.473)	1.159 (0.476)	1.151 (0.475)
Elementary school			0.900 (0.259)	0.739 (0.234)	0.905 (0.256)	0.894 (0.250)	0.888 (0.251)	0.909 (0.255)
High school			0.813 (0.271)	0.840 (0.312)	0.795 (0.265)	0.796 (0.262)	0.795 (0.265)	0.785 (0.260)
N	4770	4780	4060	3540	4070	4060	4070	4060

Table 36 (cont'd)

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). Estimates are adjusted by leaving the profession weights (i.e., not transformed weights). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 37. The Heterogeneous Effects of Principal Leadership on Leaving the Profession (Using untransformed weights)

<i>Panel A. Principal leadership*time indicators</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.829 (0.189)			
Instructional leadership		0.588 (0.209)		
Leadership related to student management			0.879 (0.190)	
Leadership related to supportive culture				0.798 (0.226)
Year 2	9320162.6*** (33455590.6)	15403344.7*** (54454987.0)	10976429.5*** (39426305.4)	15535081.1*** (54561647.6)
Year 3	3619208.1*** (13008539.5)	6253714.2*** (22062340.9)	4385502.0*** (15710154.6)	6953241.4*** (24403927.1)
Year 4	8474617.7*** (30675866.0)	11242440.4*** (39865794.5)	8477115.0*** (30581433.4)	11912164.1*** (42001084.8)
Year 5	6318610.9*** (22831481.0)	9082612.6*** (32034425.6)	7069086.8*** (25128149.4)	8977477.2*** (31583667.8)
Year 2* Leadership	0.856 (0.418)	2.440 (1.560)	1.219 (0.545)	2.392 (1.576)
Year 3* Leadership	0.637 (0.388)	0.751 (0.605)	0.458 (0.23)	1.612 (1.089)
Year 4* Leadership	0.675 (0.357)	1.258 (1.845)	1.207 (0.587)	2.032 (1.419)
N	4060	4070	4060	4070
<i>Panel B. Principal leadership*Suburban</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.787 (0.150)			
Instructional leadership		0.769 (0.256)		
Leadership related to student management			0.944 (0.181)	
Leadership related to supportive culture				0.971 (0.239)
Suburban	0.802 (0.214)	0.829 (0.218)	0.789 (0.208)	0.792 (0.207)
Suburban*Leadership	1.096 (0.275)	2.413 (1.832)	1.016 (0.451)	1.730 (0.885)
N	4060	4070	4060	4070

Table 37 (cont'd)

*Panel C. Principal leadership*Elementary school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.857 (0.161)			
Instructional leadership		0.718 (0.247)		
Leadership related to student management			0.885 (0.169)	
Leadership related to supportive culture				0.885 (0.218)
Elementary school	0.821 (0.245)	0.920 (0.261)	0.826 (0.239)	0.896 (0.256)
Elementary school*Leadership	0.604 (0.167)	1.355 (0.826)	1.816 (0.703)	2.020 (0.936)
N	4060	4070	4060	4070

*Panel D. Principal leadership*High school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.847 (0.153)			
Instructional leadership		0.750 (0.234)		
Leadership related to student management			0.977 (0.176)	
Leadership related to supportive culture				0.986 (0.235)
High school	0.852 (0.285)	0.788 (0.266)	0.817 (0.262)	0.796 (0.265)
High school*Leadership	1.363 (0.363)	0.920 (0.505)	1.241 (0.402)	1.043 (0.446)
N	4060	4070	4060	4070

*Panel E. Principal leadership*Charter school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.788 (0.150)			
Instructional leadership		0.755 (0.254)		
Leadership related to student management			0.942 (0.185)	
Leadership related to supportive culture				0.976 (0.241)
Charter school	1.135 (0.475)	1.261 (0.521)	1.187 (0.49)	1.177 (0.494)
Charter school* Leadership	1.067 (0.357)	1.95 (1.589)	1.178 (0.532)	1.083 (0.697)
N	4060	4070	4060	4070

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The variables for calculating interaction terms were centered by their grand means. Estimates

Table 37 (cont'd)

are adjusted by leaving the profession weights (i.e., not transformed weights). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix G

The Results Using No Weights

Table 38. The Influence of Principal Leadership on Leaving the School (Without weights)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.591*** (0.0369)	0.657*** (0.0532)				0.653*** (0.0495)	0.689*** (0.0637)	
General leadership (2 year lagged)		0.847* (0.0693)					0.848 (0.0744)	
Instructional leadership			0.698*** (0.0625)					0.812 (0.105)
Leadership related to student management				0.753*** (0.0422)				
Leadership related to supportive culture					0.753*** (0.0422)			
Preparation						0.893 (0.0704)	0.871 (0.0763)	0.897 (0.0712)
Alternative certificate						0.922 (0.0914)	0.952 (0.105)	0.936 (0.0927)
Induction						0.885 (0.0847)	0.889 (0.0965)	0.879 (0.0843)
Work hours						0.994 (0.00391)	1 (0.00453)	0.994 (0.00396)
Autonomy						0.847 (0.0719)	0.904 (0.086)	0.804* (0.0683)
Male						0.956 (0.0843)	1.004 (0.101)	0.936 (0.0827)
White						0.695* (0.0998)	0.693* (0.116)	0.691** (0.0969)
Union membership						0.875 (0.075)	0.887 (0.0857)	0.896 (0.0766)
Advanced degree						1.063 (0.114)	1.09 (0.131)	1.058 (0.114)
Salary(log)						0.571*** (0.0987)	0.603** (0.11)	0.585** (0.1)

Table 38 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
HQT						0.886 (0.0879)	0.898 (0.104)	0.888 (0.0882)
PD						1.041* (0.021)	1.051* (0.0236)	1.039 (0.0211)
Demanding subject						0.993 (0.0891)	1.018 (0.103)	0.995 (0.0893)
Commitment						0.603*** (0.0466)	0.691*** (0.0598)	0.562*** (0.044)
N	4880	4120	5030	5030	5000	4260	3580	4270
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
General leadership (1 year lagged)			0.686*** (0.0595)	0.740** (0.0783)				
General leadership (2 year lagged)				0.865 (0.0860)				
Instructional leadership					0.801 (0.114)			1.031 (0.171)
Leadership related to student management	0.864* (0.0598)					0.878 (0.0722)		1.024 (0.097)
Leadership related to supportive culture		0.648*** (0.0596)					0.672*** (0.0706)	0.653*** (0.084)
Preparation	0.894 (0.0708)	0.899 (0.071)	0.852 (0.0730)	0.808* (0.0767)	0.857 (0.0735)	0.856 (0.0733)	0.856 (0.0730)	0.855 (0.073)
Alternative certificate	0.925 (0.0919)	0.923 (0.0918)	0.924 (0.101)	0.951 (0.116)	0.925 (0.101)	0.927 (0.101)	0.929 (0.102)	0.930 (0.102)
Induction	0.874 (0.0836)	0.896 (0.0863)	0.929 (0.0979)	0.940 (0.111)	0.930 (0.0984)	0.926 (0.0977)	0.939 (0.0996)	0.939 (0.1)
Work hours	0.994 (0.00392)	0.994 (0.00393)	0.993 (0.00422)	0.997 (0.00485)	0.992 (0.00427)	0.993 (0.00424)	0.993 (0.00424)	0.993 (0.004)
Autonomy	0.809* (0.0686)	0.845* (0.0724)	0.880 (0.0832)	0.893 (0.0967)	0.843 (0.0797)	0.845 (0.0797)	0.883 (0.0845)	0.882 (0.085)
Male	0.917 (0.0812)	0.945 (0.0831)	0.907 (0.0891)	0.997 (0.111)	0.893 (0.0877)	0.880 (0.0864)	0.903 (0.0885)	0.904 (0.089)
White	0.696* (0.0993)	0.703* (0.0995)	0.756 (0.121)	0.777 (0.147)	0.764 (0.119)	0.766 (0.122)	0.760 (0.119)	0.763 (0.120)

Table 38 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Union membership	0.896 (0.0765)	0.883 (0.0756)	0.924 (0.0879)	0.935 (0.101)	0.944 (0.0897)	0.945 (0.0895)	0.928 (0.0884)	0.929 (0.089)
Advanced degree	1.078 (0.116)	1.068 (0.115)	1.149 (0.134)	1.152 (0.149)	1.145 (0.133)	1.166 (0.135)	1.160 (0.134)	1.165 (0.135)
Salary(log)	0.582** (0.1)	0.567*** (0.0971)	0.463*** (0.0825)	0.508*** (0.0952)	0.466*** (0.0829)	0.469*** (0.0833)	0.457*** (0.0814)	0.458*** (0.082)
HQT	0.898 (0.0893)	0.884 (0.088)	0.876 (0.0957)	0.870 (0.111)	0.876 (0.0958)	0.882 (0.0965)	0.872 (0.0954)	0.870 (0.095)
PD	1.041* (0.0211)	1.045* (0.0211)	1.040 (0.0229)	1.048 (0.0256)	1.037 (0.0229)	1.039 (0.0229)	1.043 (0.0230)	1.043 (0.023)
Demanding subject	0.982 (0.0881)	0.994 (0.0892)	0.997 (0.0980)	1.030 (0.114)	1.004 (0.0985)	0.998 (0.0979)	1.003 (0.0983)	1.002 (0.098)
Commitment	0.57*** (0.0445)	0.613*** (0.0485)	0.591*** (0.0498)	0.677*** (0.0644)	0.563*** (0.0479)	0.564*** (0.0481)	0.602*** (0.0521)	0.600*** (0.053)
School size			1.000 (0.0000933)	1.000 (0.000107)	1.000 (0.0000930)	1.000 (0.0000937)	1.000 (0.0000934)	1 (0.000934)
% of racially minority students			1.005** (0.00184)	1.006** (0.00204)	1.005** (0.00183)	1.005** (0.00184)	1.005** (0.00185)	1.005** (0.002)
School safety			0.993 (0.115)	0.987 (0.130)	0.978 (0.113)	1.000 (0.117)	1.009 (0.118)	1.007 (0.118)
Parents' involvement			0.957 (0.0691)	1.031 (0.0878)	0.948 (0.0686)	0.956 (0.0690)	0.949 (0.0688)	0.948 (0.069)
AYP status			0.796* (0.0793)	0.779* (0.0876)	0.790* (0.0784)	0.792* (0.0790)	0.807* (0.0805)	0.805* (0.081)
% of FRL			0.998 (0.00252)	0.998 (0.00287)	0.998 (0.00250)	0.998 (0.00250)	0.998 (0.00251)	0.998 (0.00251)
Suburban			0.931 (0.110)	0.848 (0.110)	0.945 (0.110)	0.937 (0.110)	0.960 (0.112)	0.962 (0.113)
Charter school			0.840 (0.187)	0.732 (0.190)	0.847 (0.189)	0.840 (0.187)	0.833 (0.186)	0.833 (0.186)

Table 38 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Elementary school			1.087 (0.146)	0.958 (0.144)	1.088 (0.146)	1.105 (0.148)	1.101 (0.148)	1.100 (0.149)
High school			0.937 (0.115)	0.869 (0.119)	0.901 (0.111)	0.900 (0.110)	0.918 (0.113)	0.921 (0.114)
N	4260	4270	3760	3160	3760	3760	3760	3760

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 39. The Heterogeneous Effects of Principal Leadership on Leaving the School (Without weights)

<i>Panel A. Principal leadership*time indicators</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.727** (0.0745)			
Instructional leadership		0.801 (0.160)		
Leadership related to student management			0.833 (0.0906)	
Leadership related to supportive culture				0.636** (0.094)
Year 2	46885.9*** (89607.8)	43485.5*** (83534.7)	41671.2*** (79273.6)	87989.7*** (170495.7)
Year 3	31825.5*** (61074.9)	25748.3*** (49665.1)	24756.7*** (47312.5)	53120.5*** (103434.0)
Year 4	34158.1*** (66296.7)	22436.8*** (43606.5)	21199.0*** (40797.7)	45387.2*** (88958.4)
Year 5	25187.2*** (48911.0)	16905.6*** (32666.6)	16056.7*** (30749.4)	34538.1*** (67414.5)
Year 2* Leadership	0.757 (0.174)	1.046 (0.471)	1.190 (0.271)	1.181 (0.392)
Year 3* Leadership	0.985 (0.285)	0.928 (0.475)	1.183 (0.297)	1.432 (0.535)
Year 4* Leadership	0.897 (0.229)	1.315 (0.691)	1.155 (0.314)	1.156 (0.461)
N	3760	3760	3760	3760
<i>Panel B. Principal leadership*Suburban</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.685*** (0.0595)			
Instructional leadership		0.797 (0.113)		
Leadership related to student management			0.878 (0.0722)	
Leadership related to supportive culture				0.655*** (0.0687)
Suburban	0.925 (0.116)	0.940 (0.111)	0.934 (0.110)	0.952 (0.113)
Suburban*Leadership	0.978 (0.166)	0.875 (0.255)	0.943 (0.155)	0.664 (0.145)
N	3760	3760	3760	3760

Table 39 (cont'd)

*Panel C. Principal leadership*Elementary school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.681*** (0.059)			
Instructional leadership		0.798 (0.114)		
Leadership related to student management			0.882 (0.0724)	
Leadership related to supportive culture				0.656*** (0.0695)
Elementary school	1.153 (0.162)	1.093 (0.146)	1.050 (0.145)	1.114 (0.149)
Elementary school*Leadership	1.232 (0.196)	1.114 (0.310)	1.402* (0.237)	1.763** (0.354)
N	3760	3760	3760	3760

*Panel D. Principal leadership*High school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.686*** (0.064)			
Instructional leadership		0.804 (0.115)		
Leadership related to student management			0.878 (0.0721)	
Leadership related to supportive culture				0.659*** (0.0695)
High school	0.942 (0.121)	0.907 (0.112)	0.896 (0.110)	0.898 (0.111)
High school*Leadership	1.018 (0.147)	1.096 (0.277)	0.957 (0.136)	0.694* (0.128)
N	3760	3760	3760	3760

*Panel E. Principal leadership*Charter school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 yr lagged)	0.687*** (0.0596)			
Instructional leadership		0.802 (0.114)		
Leadership related to student management			0.882 (0.0726)	
Leadership related to supportive culture				0.671*** (0.0705)
Charter school	0.805 (0.198)	0.804 (0.188)	0.775 (0.18)	0.844 (0.191)
Charter school* Leadership	0.882 (0.251)	0.6 (0.308)	0.639 (0.194)	1.106 (0.436)
N	3760	3760	3760	3760

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The variables for calculating interaction terms were centered by their grand means. The

Table 39 (cont'd)

coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level.

Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. *

$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 40. The Influence of Principal Leadership on Leaving the Profession (Without weights)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General leadership (1 year lagged)	0.623*** (0.0476)	0.626*** (0.0639)				0.675*** (0.063)	0.648*** (0.0744)	
General leadership (2 year lagged)		0.914 (0.105)					0.939 (0.118)	
Instructional leadership			0.683** (0.0851)					0.716 (0.125)
Leadership related to student management				0.723*** (0.055)				
Leadership related to supportive culture Preparation					0.646*** (0.0611)			
Alternative certificate						0.711*** (0.0715)	0.717** (0.0793)	0.724** (0.073)
Induction						0.952 (0.124)	1.02 (0.146)	0.984 (0.127)
Work hours						0.731** (0.0887)	0.769 (0.103)	0.729** (0.0885)
Autonomy						0.992 (0.0051)	0.993 (0.00581)	0.991 (0.00518)
Male						0.954 (0.107)	0.987 (0.12)	0.927 (0.103)
White						1.048 (0.126)	1.016 (0.135)	1.022 (0.123)
Union membership						0.774 (0.147)	0.87 (0.194)	0.766 (0.143)
Advanced Degree						0.842 (0.0955)	0.84 (0.105)	0.856 (0.0966)
Salary(log)						1.132 (0.165)	1.094 (0.176)	1.103 (0.161)
HQT						0.404*** (0.0889)	0.392*** (0.0953)	0.411*** (0.0884)
PD						0.681** (0.0881)	0.775 (0.118)	0.685** (0.0883)
						1.043 (0.0273)	1.039 (0.0298)	1.04 (0.0271)

Table 40 (cont'd)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Demanding subject						1.066 (0.125)	1.122 (0.144)	1.063 (0.124)
Commitment						0.647*** (0.0625)	0.713** (0.0752)	0.618*** (0.061)
N	5810	5060	5870	5850	5870	5090	4410	5290
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
General leadership (1 year lagged)			0.672*** (0.0735)	0.638** (0.0874)				
General leadership (2 year lagged)				0.994 (0.149)				
Instructional leadership					0.702 (0.145)			0.783 (0.181)
Leadership related to student management	0.821* (0.0758)					0.795* (0.0881)		0.836 (0.107)
Leadership related to supportive culture		0.818 (0.107)					0.809 (0.123)	0.996 (0.183)
Preparation	0.717*** (0.0721)	0.716*** (0.0719)	0.632*** (0.0727)	0.626*** (0.0794)	0.640*** (0.0735)	0.639*** (0.0735)	0.633*** (0.0727)	0.644*** (0.074)
Alternative certificate	0.972 (0.126)	0.982 (0.127)	0.836 (0.128)	0.878 (0.147)	0.836 (0.128)	0.842 (0.129)	0.844 (0.129)	0.838 (0.128)
Induction	0.72** (0.0869)	0.725** (0.0879)	0.845 (0.120)	0.880 (0.137)	0.840 (0.119)	0.836 (0.118)	0.835 (0.118)	0.844 (0.12)
Work hours	0.992 (0.00514)	0.993 (0.00509)	0.988* (0.00568)	0.989 (0.00644)	0.987* (0.00581)	0.988* (0.00579)	0.989* (0.00569)	0.987* (0.0058)
Autonomy	0.932 (0.103)	0.929 (0.104)	0.943 (0.122)	0.947 (0.135)	0.912 (0.118)	0.917 (0.118)	0.914 (0.119)	0.930 (0.121)
Male	0.993 (0.12)	1.017 (0.122)	0.996 (0.139)	0.972 (0.152)	0.972 (0.136)	0.950 (0.133)	0.966 (0.135)	0.961 (0.135)
White	0.77 (0.145)	0.776 (0.145)	0.748 (0.161)	0.839 (0.208)	0.748 (0.159)	0.750 (0.161)	0.761 (0.162)	0.741 (0.158)
Union membership	0.857 (0.0964)	0.856 (0.0966)	0.907 (0.121)	0.883 (0.130)	0.941 (0.125)	0.938 (0.124)	0.941 (0.125)	0.932 (0.124)
Advanced Degree	1.127 (0.165)	1.126 (0.164)	1.150 (0.189)	1.138 (0.209)	1.129 (0.184)	1.152 (0.189)	1.155 (0.189)	1.134 (0.185)

Table 40 (cont'd)

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Salary(log)	0.409*** (0.0884)	0.415*** (0.0895)	0.387*** (0.0862)	0.362*** (0.0909)	0.391*** (0.0860)	0.392*** (0.0867)	0.395*** (0.0869)	0.388*** (0.086)
HQT	0.693** (0.0894)	0.687** (0.0888)	0.673** (0.0985)	0.751 (0.129)	0.675** (0.0988)	0.680** (0.0995)	0.677** (0.0992)	0.675** (0.099)
PD	1.043 (0.0272)	1.039 (0.0271)	1.050 (0.0306)	1.044 (0.0329)	1.045 (0.0303)	1.049 (0.0304)	1.047 (0.0302)	1.049 (0.030)
Demanding subject	1.045 (0.122)	1.061 (0.124)	1.052 (0.139)	1.115 (0.163)	1.055 (0.139)	1.042 (0.138)	1.051 (0.139)	1.050 (0.139)
Commitment	0.623*** (0.0606)	0.619*** (0.0628)	0.615*** (0.0700)	0.691** (0.0887)	0.601*** (0.0697)	0.604*** (0.0687)	0.598*** (0.0707)	0.622*** (0.075)
School size			1.000 (0.000135)	1.000 (0.000152)	1.000 (0.000134)	1.000 (0.000135)	1.000 (0.000135)	1 (0.000134)
% of racially minority students			1.004 (0.00247)	1.006* (0.00262)	1.005* (0.00243)	1.004 (0.00246)	1.005 (0.00246)	1.004 (0.0025)
School safety			1.208 (0.185)	1.324 (0.225)	1.150 (0.179)	1.186 (0.186)	1.159 (0.180)	1.183 (0.185)
Parents' involvement			0.998 (0.0992)	1.037 (0.118)	0.998 (0.0988)	1.007 (0.0999)	0.994 (0.0988)	1.006 (0.1001)
AYP status			0.915 (0.127)	0.924 (0.143)	0.926 (0.128)	0.931 (0.129)	0.922 (0.128)	0.943 (0.131)
% of FRL			0.997 (0.00328)	0.997 (0.00365)	0.997 (0.00328)	0.996 (0.00327)	0.997 (0.00329)	0.996 (0.0032)
Suburban			0.900 (0.152)	0.819 (0.153)	0.915 (0.153)	0.898 (0.150)	0.910 (0.152)	0.908 (0.152)
Charter school			1.144 (0.309)	1.085 (0.333)	1.210 (0.324)	1.192 (0.316)	1.202 (0.322)	1.193 (0.318)
Elementary school			0.895 (0.172)	0.760 (0.162)	0.893 (0.171)	0.912 (0.174)	0.900 (0.172)	0.910 (0.174)
High school			1.293 (0.220)	1.254 (0.232)	1.229 (0.209)	1.219 (0.208)	1.247 (0.212)	1.216 (0.208)
N	5090	5090	4320	3730	4330	4320	4330	4320

Table 40 (cont'd)

Note. Time dummies were included in all the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 41. The Heterogeneous Effects of Principal Leadership on Leaving Profession (Without weights)

<i>Panel A. Principal leadership*time indicators</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.660*** (0.0790)			
Instructional leadership		0.730 (0.176)		
Leadership related to student management			0.826 (0.108)	
Leadership related to supportive culture				0.880 (0.163)
Year 2	147433.4*** (350944.0)	104463.6*** (244924.1)	92103.7*** (218410.3)	77058.3*** (183118.1)
Year 3	107426.0*** (258277.0)	65702.7*** (155041.1)	60027.9*** (143252.9)	50944.9*** (121904.6)
Year 4	174647.3*** (422032.6)	87066.8*** (207000.2)	78567.1*** (188870.0)	66351.7*** (159836.9)
Year 5	148587.5*** (360376.4)	71961.3*** (170343.1)	63568.7*** (152376.4)	53507.5*** (128624.6)
Year 2* Leadership	1.038 (0.287)	0.869 (0.423)	0.859 (0.237)	0.726 (0.283)
Year 3* Leadership	1.238 (0.401)	0.555 (0.327)	0.782 (0.249)	0.712 (0.311)
Year 4* Leadership	1.224 (0.34)	0.784 (0.434)	0.975 (0.308)	1.043 (0.458)
N	4320	4330	4320	4330
<i>Panel B. Principal leadership*Suburban</i>				
	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.677*** (0.0735)			
Instructional leadership		0.699 (0.145)		
Leadership related to student management			0.795* (0.0882)	
Leadership related to supportive culture				0.820 (0.125)
Suburban	0.945 (0.163)	0.906 (0.156)	0.901 (0.153)	0.922 (0.155)
Suburban*Leadership	1.219 (0.232)	0.883 (0.363)	1.027 (0.233)	1.292 (0.410)
N	4320	4330	4320	4330

Table 41 (cont'd)

*Panel C. Principal leadership*Elementary school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.672*** (0.0735)			
Instructional leadership		0.699 (0.145)		
Leadership related to student management			0.801* (0.089)	
Leadership related to supportive culture				0.809 (0.124)
Elementary school	0.913 (0.181)	0.912 (0.175)	0.887 (0.173)	0.924 (0.178)
Elementary school*Leadership	1.072 (0.207)	1.291 (0.540)	1.463 (0.352)	2.093* (0.639)
N	4320	4330	4320	4330

*Panel D. Principal leadership*High school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.675*** (0.0745)			
Instructional leadership		0.702 (0.146)		
Leadership related to student management			0.796* (0.0879)	
Leadership related to supportive culture				0.804 (0.123)
High school	1.315 (0.228)	1.231 (0.212)	1.242 (0.211)	1.228 (0.208)
High school*Leadership	1.076 (0.186)	1.023 (0.373)	1.145 (0.227)	0.829 (0.224)
N	4320	4330	4320	4330

*Panel E. Principal leadership*Charter school*

	Model 1	Model 2	Model 3	Model 4
General leadership (1 year lagged)	0.671*** (0.0738)			
Instructional leadership		0.698 (0.146)		
Leadership related to student management			0.803 (0.0899)	
Leadership related to supportive culture				0.802 (0.123)
Charter school	1.165 (0.325)	1.235 (0.343)	1.103 (0.316)	1.252 (0.340)
Charter school* Leadership	1.064 (0.298)	1.174 (0.716)	0.736 (0.241)	1.318 (0.585)
N	4320	4330	4320	4330

Note. Time dummies and all teacher level- and school level- control variables were included in the models. General principal leadership, salary, HQT status, percentage of free and reduced lunch eligible

Table 41 (cont'd)

students, charter school, and suburban were included as time-variant variables (i.e., lagged variables were included). The variables for calculating interaction terms were centered by their grand means. The coefficients are reported as odds ratios. The standard errors were clustered at the individual teacher level. Sample sizes were rounded in the closest ten's place according to NCES nondisclosure regulations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

REFERENCES

REFERENCES

- Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, 39(1), 54-76.
- Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago public schools*. Chicago, IL: Consortium on Chicago School Research.
- Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Teacher churning: Reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis*, 39(1), 3-30.
- Berliner, D.C. (1988). Implications of studies in pedagogy for teacher education and evaluation. In *New directions for teacher assessment*. Princeton, NJ: Educational Testing Service.
- Borman, G. D., & Dowling, N. M. (2008). Teacher attrition and retention: A meta-analytic and narrative review of the research. *Review of Educational Research*, 78(3), 367-409.
- Burns, S., Wang, X., & Henning, A. (2011). NCES handbook of survey methods. NCES 2011-609. *National Center for Education Statistics*.
- Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, 48(2), 303-333.
- Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, 27(4), 793-818.
- Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the short careers of high-achieving teachers in schools with low-performing students. *The American Economic Review*, 95(2), 166-171.
- Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers. *Journal of Labor Economics*, 31(1), 83-117.
- Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage Foundation.
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6), 673-682.
- Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management*, 23(2), 251-271.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

- Darling-Hammond, L., & Sykes, G. (2003). Wanted: A national teacher supply policy for education: The right way to meet the “Highly Qualified Teacher” challenge. *Education Policy Analysis Archives*, 11(33). Retrieved from <http://epaa.asu.edu/epaa/v11n33>
- DeAngelis, K. J. & Presley, J. B. (2007). *Leaving schools or leaving the profession: Setting Illinois' record straight on new teacher attrition* (IERC 2007-1). Edwardsville, IL: Illinois Education Research Council.
- Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research*, 29(2), 147-194.
- Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin’s causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, 35(4), 437-460.
- Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? *Education Evaluation and Policy Analysis*, 30(1), 3–30.
- Fuller, F. F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal*, 6(2), 207-226.
- Goldhaber, D., & Hansen, M. (2009). National Board Certification and teachers' career paths: Does NBPTS certification influence how long teachers remain in the profession and where they teach? *Education Finance and Policy*, 4(3), 229-262.
- Grissom, J. A. (2011). Can good principals keep teachers in disadvantaged schools? Linking principal effectiveness to teacher satisfaction and turnover in hard-to-staff environments. *Teachers College Record*, 113(11), 2552-2585.
- Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher*, 42(8), 433–444.
- Guin, K. (2004). Chronic teacher turnover in urban elementary schools. *Educational Policy Analysis Archives*, 12(42), 1–25.
- Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, 33(3), 329–352.
- Hallinger, P. (2005). Instructional leadership and the school principal: A passing fancy that refuses to fade away. *Leadership and Policy in Schools*, 4(3), 221–239.
- Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources*, 39(2), 326–354.
- Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher*, 40(6), 271-280.
- Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of*

- time-to-event data*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, 38(3), 499-534.
- Ingersoll, R. M., & May, H. (2012). The magnitude, destinations, and determinants of mathematics and science teacher turnover. *Educational Evaluation and Policy Analysis*, 34(4), 435-464.
- Ingersoll, R., Merrill, L., & Stuckey, D. (2014). Seven trends: The transformation of the teaching force. *CPRE Research Reports*. Retrieved from http://repository.upenn.edu/cpre_researchreports/79
- Johnson, S. M., & Birkeland, S. E. (2003). Pursuing a “sense of success”: New teachers explain their career decisions. *American Educational Research Journal*, 40(3), 581-617.
- Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers’ working conditions on their professional satisfaction and their students’ achievement. *Teachers College Record*, 114(10), 1-39.
- Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, 62(2), 129-169.
- Kaiser, A. (2011). *Beginning Teacher Attrition and Mobility: Results From the First Through Third Waves of the 2007-08 Beginning Teacher Longitudinal Study* (NCES 2011-318). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from <http://nces.ed.gov/pubsearch>.
- Kelly, S. (2004). An event history analysis of teacher attrition: Salary, teacher tracking, and socially disadvantaged schools. *The Journal of Experimental Education*, 72(3), 195-220.
- Kelly, S., & Northrop, L. (2015). Early career outcomes for the “best and the brightest”: selectivity, satisfaction, and attrition in the beginning teacher longitudinal survey. *American Educational Research Journal*, 52(4), 624-656.
- Ladd, H. F. (2011). Teachers’ perceptions of their working conditions how predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis*, 33(2), 235-261.
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, 24(1), 37-62.
- Lesik, S. A. (2007). Do developmental mathematics programs have a causal impact on student retention? An application of discrete-time survival and regression-discontinuity analysis. *Research in Higher Education*, 48(5), 583-608.
- Liu, X. (2012). *Survival analysis: Models and applications*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Loeb, S., Darling-Hammond, L., & Luczak, J. (2005). How teaching conditions predict teacher turnover in California schools. *Peabody Journal of Education*, 80(3), 44-70.
- Louis, K. S., Marks, H. M., & Kruse, S. (1996). Teachers’ professional community in restructuring schools. *American Educational Research Journal*, 33(4), 757-798.

- Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly*, 39(3), 370–397.
- Murphy, J. (1988). Methodological, measurement, and conceptual problems in the study of instructional leadership. *Educational Evaluation and Policy Analysis*, 10(2), 117–139.
- Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, 40(3), 355–371.
- Plank, S. B., DeLuca, S., & Estacion, A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education*, 81(4), 345–370.
- Pogodzinski, B., Youngs, P., Frank, K. A., & Belman, D. (2012). Administrative climate and novices' intent to remain teaching. *The Elementary School Journal*, 113(2), 252–275.
- Redding, C., & Smith, T. M. (2016). Easy in, easy out: Are alternatively certified teachers turning over at increased rates? *American Educational Research Journal*, 53(4), 1086–1125.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.
- Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, 50(1), 4–36.
- Rorrer, A. K., & Skrla, L. (2005). Leaders as policy mediators: The reconceptualization of accountability. *Theory into Practice*, 44(1), 53–62.
- Scafidi, B., Sjoquist, D. L., & Stinebrickner, T. R. (2007). Race, poverty, and teacher mobility. *Economics of Education Review*, 26(2), 145–159.
- Schultz, B. K., Evans, S. W., & Serpell, Z. N. (2009). Preventing failure among middle school students with attention deficit hyperactivity disorder: A survival analysis. *School Psychology Review*, 38(1), 14–27.
- Simon, N. S., & Johnson, S. M. (2013). Teacher turnover in high-poverty schools: What we know and can do. *Teachers College Record*, 117(3), 1–36.
- Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, 18(2), 155–195.
- Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal*, 41(3), 681–714.
- Spillane, J. P., Halverson, R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher*, 30(3), 23–28.
- Supovitz, J., Sirinides, P., & May, H. (2009). How principals and peers influence teaching and learning. *Educational Administration Quarterly*, 46(1), 31–56.

- Prater, D. L., Bermudez, A. B., & Owens, E. (1997). Examining parental involvement in rural, urban, and suburban schools. *Journal of Research in Rural Education*, 13(1), 72-75.
- Youngs, P., Kwak, H. S., & Pogodzinski, B. (2015). How middle school principals can affect beginning teachers' experiences. *Journal of School Leadership*, 25(1). 157-189.