# The Transfer of Educational Policies and Practices Across International Borders

By

Olivia G. Carr

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Leadership and Policy Studies

May 8, 2020
Nashville, Tennessee

Approved:

Carolyn J. Heinrich, Ph.D.

Xiu Cravens, Ph.D.

Brian L. Heuser, Ed.D.

Kathryn Anderson, Ph.D.

To my parents, Gregg and Susan

and

To my husband, Toren

## ACKNOWLEDGEMENTS

I would like to thank those who supported me throughout graduate school and this process of writing a dissertation. I am grateful to my dissertation committee for their honesty, positivity, and support. They provided extremely valuable feedback on my dissertation at every stage of the process and have helped me navigate graduate school and my future career. They have been caring role models and mentors to me for my entire time at Vanderbilt.

I am forever indebted to my friends and family, who supported me endlessly through this process. My parents, Dr. Susan Murray and Dr. Gregg Murray, and beloved late grandfather, Dr. Robert Martin, inspired me to pursue a Ph.D. I know I am here today due to the love and guidance from them and my entire family. Lastly, I wish to thank my husband, Toren, for his love and steadiness and for always knowing how to make me smile.

# PREFACE

The title of this dissertation refers to educational transfer. There are a variety of terms that relate to transfer, including lesson drawing (Rose, 1991) and policy borrowing (Phillips, 2009), but I use the term "transfer" to broadly describe "the movement of ideas, structures and practices in education policy, from one time and place to another" (Perry & Tor, 2009, p. 510). It describes the diffusion, adoption, adaptation, convergence, and possible rejection of policies and practices. It is believed that policy transfer is happening more often due to modern communication and transportation systems, competition between countries, interdependent systems and institutions, and the popularity of evidence-based policymaking (McDonald, 2012; Minkman, van Buuren, & Bekkers, 2018).

One of the major theories as to why educational transfer happens is institutional isomorphism (DiMaggio & Powell, 1983; Meyer & Rowan, 1977). According to Meyer and Rowan's theory, institutions are driven by the desire to maintain legitimacy among other institutions, rather than the desire to increase efficiency. Because of this, institutions will adopt concepts or practices that have been deemed legitimate by external bodies, which improves the institutions' chances of being successful and surviving over time. Nation-states display a strong tendency toward isomorphism in their policies and structures, either through their own desires to be considered legitimate or from the imposition of citizens internally or external actors like the World Bank (Meyer, Boli, Thomas, & Ramirez, 1997).

According to DiMaggio and Powell (1983), there are three mechanisms through which isomorphic change happens: coercive, mimetic, and normative. Coercive isomorphism is when powerful institutions apply pressure to organizations to force or persuade the desired changed into effect. Mimetic isomorphism is a softer approach by which organizations decide to model themselves off of others that they perceive to be legitimate and/or successful. Normative isomorphism is change inspired by a desire for professionalization, where the expectation is that a profession will gain legitimacy, autonomy, and/or efficiency if the related social actors and institutions converge on professional features such as qualifications for entry into the profession and job titles.

Educational transfer by way of isomorphism is the inspiration for this dissertation. In the course of three papers, I will be studying three aspects of the policymaking process: adoption, adaptation, and evaluation. My first paper is called "Promoting Priorities: Explaining the Adoption of Compulsory Education Laws in Africa," and I use an event history analysis to examine the diffusion of lower secondary compulsory schooling laws throughout Africa. The current theory on compulsory schooling law diffusion has been developed almost exclusively by studying patterns of adoption in the Western world. In this paper, I test reasons for diffusion that are prominent in this Western literature and add additional predictors that are likely to better represent educational governance in Africa. The main finding from this paper is that lower secondary compulsory schooling law adoption in Africa can be predicted primarily through common linguistic and historical ties between countries, rather than through other variables that are theorized to be important from prior literature.

My second paper is called "Teaching Without Boundaries: Adapting Collaborative Inquiry Cycles to the American Context." For this paper, I qualitatively study three Tennessee schools that were pilot sites for a teacher collaboration model called Teacher Peer Excellence Groups (TPEG), which is a version of the Japanese or Chinese lesson study. Through interviews and observations, I examine the enabling conditions that support TPEG cycles, how TPEG has adapted to the local contexts to become sustainable collaborative practices, and whether these adaptations have resulted in professional knowledge-building as intended from TPEG. Results reveal the particular importance of instructional leadership for the sustainability of TPEG and show how teachers maintain many steps of TPEG but replace the observation step with other practices to promote the transition from a practitioner knowledge base to a professional knowledge base.

The third paper in this dissertation is called "The Use of Student Assessments to Evaluate Teachers: An International, Longitudinal Study." For this paper, I use data from four waves of the Programme for International Study Assessment and a country fixed effects model to test the expectation that incorporating student assessment data into teacher evaluations improves student learning. To overcome problems with selection bias, I aggregate the use of this type of evaluation system to the country-year level. The results indicate that using student assessment data in teacher evaluations does increase student learning in mathematics (but not reading), though this effect is more prevalent in country-years with a low GDP per capita, and the effects are reversed for country-years with a high GDP per capita.

My intent for this dissertation is to influence the conversation about international educational transfer. Through these three papers, I extend the compulsory schooling law adoption theory to a new context, provide insight on the sustainability and adaptation of a promising teacher collaboration model, and test a common assumption about a type of teacher evaluation. These papers also reflect my research interests, provide interesting directions for future research, and showcase my range of methodological expertise. I hope readers can appreciate the value of this research and also the time, effort, and love put into it by myself and everyone who supported me in this process.

## REFERENCES

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, *48*(2), 147–160. https://doi.org/10.1016/S0742-3322(00)17011-1

McDonald, L. (2012). Educational Transfer to Developing Countries: Policy and Skill Facilitation. *Procedia - Social and Behavioral Sciences*, *69*, 1817–1826. https://doi.org/10.1016/j.sbspro.2012.12.132

Meyer, J., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *The American Journal of Sociology*, *83*(2), 340–363. https://doi.org/10.17323/1726-3247-2011-1-43-67

Meyer, J. W., Boli, J., Thomas, G. M., & Ramirez, F. O. (1997). World Society and the Nation-State. *The American Journal of Sociology*, *103*(1), 144–181. Retrieved from http://www.jstor.org/stable/2782801

Minkman, E. (Ellen), van Buuren, M. W. (Arwin., & Bekkers, V. J. J. M. (Victor. (2018). Policy transfer routes: an evidence-based conceptual model to explain policy adoption. *Policy Studies*, *39*(2), 222–250. https://doi.org/10.1080/01442872.2018.1451503

Perry, L. B., & Tor, G. (2009). Understanding educational transfer: Theoretical perspectives and conceptual frameworks. *Prospects*, *38*(4), 509–526. https://doi.org/10.1007/s11125-009-9092-3

Phillips, D. (2009). Aspects of Educational Transfer. In R. Cowen & A. M. Kazamias (Eds.), *Second International Handbook of Comparative Education* (pp. 1061–1077). Springer.

Rose, R. (1991). What Is Lesson-Drawing? *Journal of Public Policy*, *11*(1), 3–30.

# TABLE OF CONTENTS

# LIST OF TABLES

Chapter 1

Table Page

Chapter 2

Table Page

Chapter 3

Table Page

# LIST OF FIGURES

Chapter 1

Figure Page

Chapter 2

Figure Page

Chapter 3

Figure Page

# CHAPTER 1

## PROMOTING PRIORITIES: EXPLAINING THE ADOPTION OF COMPULSORY EDUCATION LAWS IN AFRICA

According to the OECD (1983), compulsory schooling encompasses the years in which all typically-developing children must receive a formal education. The United Nations' Declaration of Human Rights (Article 26) emphasizes the right to a free elementary education, and by 1985, 80 percent of countries in the world had compulsory education (Ramirez, 1989). By formalizing the state's role in its provision, compulsory schooling has been lauded as a state and societal recognition of the importance of education in modern and democratic societies (Nawaz & Tanveer, 1975). Along with child labor laws, compulsory schooling laws (CSLs) seek to prompt parents to voluntarily comply by keeping their children out of the workforce and in schools, though historically, the laws included many exceptions and were poorly enforced (Nawaz & Tanveer, 1975; Oreopoulos, 2009).

The potential benefits of schooling are called the returns to education, and they encompass a multitude of outcomes of interest, including earnings, health, and civic engagement. These outcomes are then divided into private and public returns to education. Private returns to education can be measured in terms of individual earnings or other benefits gained from attending an additional year of schooling. In addition to the economic benefits that are commonly cited, other returns to education, such as becoming an effective voice in the household and community; staying free, informed, and healthy; and becoming socially, economically, and geographically mobile, are important components of the private returns to education (Hill & King, 1993; Lloyd, Kaufman, & Hewett, 2000; Nawaz & Tanveer, 1975; Psacharopoulos & Patrinos, 2018).

The second type of returns are public or social returns, which can be measured as the increase in total earnings or other benefits to society from one additional year of average schooling in a population. The social returns to education are the primary reason why the public funds and mandates education in so much of the world. The mass accumulation of human capital and other private and social benefits in a country should lead to a decrease in crime and disorder and increased economic activity (Astiz, Wiseman, & Baker, 2002; Eckstein, 1994; Fuller & Rubinson, 1992). Thus, increased educational levels of citizens in a mass education system, along with other efforts for economic and social change, such as increasing the quality of education and investments in economic sectors, should advance societies into higher levels of global economic competitiveness (Acemoglu & Angrist, 2000; Akkari, 2004; Astiz et al., 2002; Deininger, 2003; Eckstein, 1994; Fataar, 1997).

State interest in mass education systems have centered on this international competitiveness, the state's own perceived institutional character, and institutional isomorphism, and every developed country has CSLs, regardless of their political or philosophical stances (Astiz et al., 2002; Nawaz & Tanveer, 1975; Ramirez & Boli, 1987). Africa, however, has had historically lower levels of development of mass education systems compared to the rest of the world (WCEFA, 1990).

This paper examines the spread of CSLs in Africa, which can be seen as a means of symbolic state support for the education of all children (Meyer, Ramirez, & Soysal, 1992), whether or not the CSLs are actually enforced. In particular, I examine what conditions predict the adoption of CSLs for lower secondary school throughout Africa. The enactment of CSLs does not necessarily mean they will be enforced, but the laws do show a societal (government and/or citizen) interest in higher levels of education for all children and open the door for their enforcement. To study this question, I use an event history analysis (EHA), which combines internal country characteristics with diffusion from nearby countries (Berry & Berry, 1990), with data primarily from the World Bank.

## Review of the Literature

### *Expansion of Formal School*

Two primary assumptions underpin the theory of action behind CSLs. The first is that CSLs increase educational attainment. This assumption relies on compliance with the laws, or more precisely, the feasibility of CSL enforcement, which often is not high, due to problems with financing or the priorities of policymakers, schools, and enforcement officers (Fyfe, 2005; Landes & Solmon, 1972; Murtin & Viarengo, 2011; Nawaz & Tanveer, 1975; Oreopoulos, 2009; Richardson, 1980). However, CSLs, along with state and citizen support and some level of enforcement or complementary laws, such as child labor laws, do seem to increase educational attainment.

The second assumption is that additional years of formal schooling benefit individuals, particularly for students least likely to attend school before the enactment of the CSLs. The nation-building driver behind many CSLs points to societal benefits of mass education. For example, Canadian CSLs were instated after concerns that poor attendance jeopardized societal goals for mass public education (Oreopoulos, 2006b). However, the literature also has a plethora of studies that show the individual benefits of additional years of education. These benefits can follow two distinct hypotheses. The human capital hypothesis assumes that CSLs will only affect the enrollment of those who otherwise would have dropped out below the minimum allowable age. The educational sorting hypothesis suggests that CSLs should affect the educational attainment even of students not directly affected by the law, meaning those who already were going to attend until the compulsory attendance age, as well (Lang & Kropp, 1986). Lang and Kropp (1986), for instance, find support for the sorting hypothesis in the United States because CSLs increased the enrollment rates of children in age groups in the U.S. that the laws did not directly affect.

Supporters of CSLs often frame the discussion around human rights, particularly equality of educational opportunity. Using a social justice framework, the primary reason why it is so important for a central government, rather than local governments or families, to mandate education is that this is the only way to ensure access for all students, regardless of region or identity group (Fyfe, 2005; Hill, Baxen, Craig, Namakula, & Sayed, 2012). Fyfe (2005) argues that education is necessary for the welfare of individuals and society, so education must be free and compulsory. The National Education Co-ordinating Committee (1989) agrees, saying that education is a basic human right and must be free and compulsory for every child.

Researchers have typically found that some degree of minimum compulsory schooling increases economic output, reduces income inequality, and provides social benefits, such as keeping people healthy and informed (Eckstein, 1994; Hill & King, 1993). To be more specific, one study from the U.S. finds that one additional year of schooling is associated with a 7 percent increase in average (social) wages. This is on top of an additional 7 percent increase in private returns when using an ordinary least squares model or a 1-2 percent increase in public returns when using instrumental variables (Acemoglu & Angrist, 2000).

Philip Oreopoulos has conducted numerous studies regarding the two key CSL assumptions internationally. He studied changes in school leaving ages in Britain and Northern Ireland and found large benefits from compulsory schooling, even if the laws only impacted a small number of the students who were exposed to the law. He found that essentially everyone who would have left school at the minimum allowed age of 14 instead left school at age 15, the new minimum age, thus increasing their level of educational attainment. From this change in attainment, he identified a number of benefits to the students, including increased earnings, health, employment, and other labor market outcomes (Oreopoulos, 2006a). Oreopoulos (2006b, 2007) suggests that some students or families ignore or heavily discount these potential benefits of additional schooling when deciding to drop out.

He explores similar benefits using samples from the U.S. and Canada, finding that additional mandated years of education led to significantly increased lifetime wealth, health, and employment, including employment outside manual labor, happiness, and proficiency in multiple languages (Oreopoulos, 2006b, 2007, 2009). He concludes that would-be dropouts might consider even higher levels of degree attainment, such as higher education, if they are mandated to remain in school longer (Oreopoulos, 2009). Oreopoulos, Page, and Stevens (2006) also find some evidence of intergenerational benefits from additional schooling in the U.S.: a one-year increase in the education of a parent decreases the likelihood that the child will repeat a grade by 2 to 4 percentage points.

Brunello, Fort, and Weber (2009) use minimum school leaving age laws (another term for CSLs) as exogenous assignment of education to individuals. They find that variation in these CSLs among twelve European countries significantly affects educational attainment. The effect is particularly strong among students from low quantiles of ability. Specifically, an additional year of compulsory schooling results in 0.30 to 0.40 years of additional education for those in lower quantiles of ability and 0.10 years of additional education for the rest of the population. The authors also find that this additional education reduces income inequality.

Using changes in CSLs in the U.S. and Norway, Black, Devereux, and Salvanes (2008) demonstrate that CSLs can have beneficial spillover effects for teenagers beyond the traditional measures of school attainment and achievement. Their study examines CSLs and teenage births to see if additional schooling encourages delayed childbearing. They find that increasing compulsory schooling leads to lower rates of teenage childbearing in both countries. A valuable part of this work is that the authors explain two mechanisms by which this relationship likely exists. First, there is the incarceration effect, meaning that compulsory schooling reduces the time available for teenagers to participate in risky behavior. This incarceration or custodial function of compulsory schooling extends logically to the lower grades as well. Even without academic benefits, compulsory school functions to keep children and teenagers busy and away from dangerous or risky situations. The other mechanism the authors address is the human capital effect, which suggests that increases to current and expected future human capital could change fertility decisions. This presumably also applies to decisions about other aspects of a student’s life, including decisions about further educational attainment, employment, and partner selection. Black and colleagues conclude that the reduction they find in teenage childbearing likely follows from both of those mechanisms.

Not all existing research, however, finds positive effects of compulsory schooling. For instance, Pischke and von Wachter (2008) use the Qualification and Career Survey and the Micro Census in Germany to study returns to education as CSLs changed from 1948 to 1970. They find no additional returns to more years of compulsory schooling, but they believe this is due to German students’ mastery of basic academic skills early in their educational career. An additional year of more advanced academic skills would thus not substantially affect these students’ labor market outcomes. This study provides insight on the importance of both quality and relevance of formal education to ensure that increased quantity of schooling leads to improved outcomes. Examining endogeneity and CSLs in the United States from 1940 to 1960, Edwards (1978) found that CSLs were not in themselves effective in increasing the enrollment of teenagers, except for 16-year-old males in 1960.

*Education in Africa*

Delegates from 63 countries affirmed the goal of universal primary education in Africa at an education conference of African states in Addis Ababa, Ethiopia, in 1961 (UNESCO, 1961). Primary enrollments soared on the continent throughout the 1960s and 1970s, though this was accompanied by vast differences in the quality of education among individual countries. Primary enrollment growth stagnated in the 1980s due to economic crisis and debt restructuring, which led to higher school fees for families and lower school quality. Most sub-Saharan African countries announced programs for universal primary education at the World Conference on Education for All in Jomtien, Thailand, in 1990, but by 1999, there were still 120 million children of primary-school age who were not in school, most of whom lived in South Asia and sub-Saharan Africa (ADB, 2002; WCEFA, 1990). The Millennium Development Goals of 2000 again sought to give all children a full course in primary schooling internationally, with some degree of success in enrollment but not necessarily retention (Hill et al., 2012; Lewin & Sabates, 2011). Several countries have since extended their goals to include lower secondary education (Lewin & Sabates, 2011), the focus of this paper.

While net enrollment rates increased dramatically over the past half century, particularly in sub-Saharan Africa, completion rates still stagnated or declined in many countries, and rapid population expansion strained the supply and quality of education and led to declining enrollments and problems with infrastructure, teacher supply, and overage children (Fyfe, 2005; Kifle et al., 2002; Lewin & Sabates, 2011; Lloyd & Blanc, 1996; Schultz, 1999; *The Millennium Development Goals Report*, 2015; Tikly, 2001). Many populations on the continent regularly deal with weak school networks, assumptions (correct or incorrect) about the quality and appropriateness of formal schooling, health and HIV/AIDS, high proportions of child laborers, national debt, large nomadic populations, and regional conflict (Carr-Hill, Eshete, Sedel, & de Souza, 2005; Fyfe, 2005). Access, attendance, and age of attendance (whether the child is attending at the appropriate grade level) are strongly associated with wealth, rural/urban status, and gender in many (particularly Francophone) countries (Akkari, 2004; Carr-Hill et al., 2005; Lewin & Sabates, 2011; Michaelowa, 2001). This slow and uneven growth of school enrollment and attendance identifies CSLs in Africa as one path to push the societal and governmental focus toward improving educational systems.

Bennell (1996) makes the argument that traditional theories of education policy do not always work well when considering sub-Saharan Africa, for a variety of theoretical and empirical reasons. He explains that the work of George Psacharopoulos and others on universal theories of educational rates of return – in particular, that private and social rates of return are highest for primary education and that private rates of return to higher education are typically much higher than social returns – guided public policy and aide efforts for decades. Instead, Bennell says, we should get rid of continent-wide rates of return rules and instead look at country-by-country variations to inform investment priorities. The report does suggest, however, that primary education might not provide as high private or social returns as researchers and practitioners expect for Africa; instead secondary school provides the highest returns in some situations.

My study also highlights the difficulty with rates of return research in the region. First, comprehensive and high-quality surveys are less available, which has limited the number of studies that can estimate the rates of return in the region. The Ugandan program of Universal Primary Education (UPE) in 1997, though, provided an evaluation of a policy that banned primary enrollment fees. This ban was associated with a dramatic increase in primary school attendance, particularly for girls, who had faced substantial direct and indirect barriers to enrollment; the ban was also correlated with income, gender, and region enrollment inequalities (Deininger, 2003).

Although there is mixed evidence in the broader literature that compulsory schooling has a positive effect on behavior and achievement, the literature does tend to support the notion that increasing schooling, including through enforced CSLs, is good for individuals and society, particularly in situations with already lower levels of education, such as sub-Saharan Africa.

### *Diffusion of Compulsory Schooling Laws*

The majority of American children in the mid-1800s attended public schools voluntarily when CSLs began to roll out, as was usual in Western education systems (Bandiera, Mohnen, & Rasul, 2015; Fyfe, 2005). To be successful, CSLs were likely to be instated around the same time as birth registration laws and child labor laws. Together, these laws effectively removed children from the labor force as the demand for child labor decreased and ideologies changed, prioritizing adult men as the primary income generators and children as deserving a childhood (Fyfe, 2005; Weiner, 1991). The laws as a whole intended to reach the few children who were not yet served in school systems, either to represent the norms of the state in favor of universal education or to respond to a sense that people were not obtaining enough general or civic education (Bandiera et al., 2015; Fyfe, 2005; Lang & Kropp, 1986).

CSLs came into effect in many Western countries primarily as a mechanism for nation-building and the dissemination of civic values. Pietism in the 1600s invigorated the compulsory schooling movement in Central Europe as a way to expand state authority, maintain social control, evolve the bureaucratic institutions, and allow for nations to “catch up” with more advanced neighbors (Van Horn Melton, 1988). This story accompanies the implementation of CSLs in almost every Western nation. CSLs were instated to integrate citizens into the nation-state; strengthen group norms; invest in human capital, civic nature, and moral character; and advance other political, economic, and cultural developments for globalization and competition (Bandiera et al., 2015; Meyer et al., 1992; Murtin & Viarengo, 2011; Ramirez & Boli, 1987; Richardson, 1980; Tyack, 1976).

Murtin and Viarengo (2011) examine the expansion and convergence of CSLs in fifteen Western European countries from 1950 to 2000. They find that lower starting minimum leaving ages predict earlier and larger changes in CSLs. They also find that countries that are more open internationally enact more years of compulsory schooling, suggesting that the desire for globalization is driving the public investment in education.

Meyer, Ramirez, and Soysal (1992) use enrollment data from 1870 to 1980 in over 120 countries to perform pooled panel regressions on educational enrollment. They find that mass education systems expanded steadily before the 1940s with a large expansion after World War II. As for national characteristics, they find that only national independence is a statistically significant predictor, whereas urbanization, percentage white, Christianity, and compulsory education rules are not. The authors take this to mean that as soon as a country forms a mass education system, actual enrollment growth does not vary significantly with the included predictive characteristics of the country.

For the United States, Bandiera et al. (2015) argue that CSLs spread among states because of a desire to instill civic values in the diverse groups of immigrants who came to the country with a low demand for American common schooling. The authors come to this conclusion after finding that voters passed CSLs earlier in states with higher proportions of European immigrants from countries without compulsory schooling. Their results were robust to accounting for the endogenous settlement localities of immigrants. Richardson (1980) also highlights the importance in the U.S. of societal occupational expansion from agriculture to other fields and the capacity of local officials to enforce the laws in predicting the passing of CSLs.

No research that I know of has looked specifically at the diffusion of CSLs throughout Africa, but existing research on education in the region can provide insight on how CSLs might diffuse differently than in the West. Colonialism was a mechanism for the spread of modern schooling in Africa as well as a point of global networks for other economic, political, and cultural developments (Tikly, 2001). After independence, some African governments deliberately tried to imitate European advancements in technology and education (Akkari, 2004). However, there was great local resistance to global forces, and colonies were less closely linked to world centers, particularly sub-Saharan colonies (Meyer et al., 1992; Tikly, 2001). Free and public education was a cornerstone policy in every North African country since independence, but localities struggled with universal primary education alongside rapid population expansion and large rural and nomadic groups (Akkari, 2004). Therefore, colonial or other historical ties will likely be an important predictor of the diffusion of CSLs in Africa.

African democratic transitions across the continent in the early 1990s are an area of focus for researchers to examine how ‘weak institutions,’ with few restrictions or accountability mechanisms for politicians, deal with policy, policy enforcement, and public spending. Stasavage (2005) argues that African governments are willing to spend more on education if that helps obtain electoral majorities in elections. He finds through a game theory model that democratically-elected African governments spend more on primary education, which is more important to rural groups (the majority of citizens in most African countries) than to urban groups, than do their non-democratic counterparts. This finding shows that democratically-elected African government officials tend to respond to constituent concerns about education, suggesting they would be willing to support CSLs should citizens demand it. This suggests that governance might be important in the diffusion of these CSLs.

Violence and political unrest are likely to play a role in CSL diffusion in Africa, as many countries in Africa experienced relatively high rates of violence and war since independence compared to countries in other regions of the world (Elbadawi & Sambanis, 2000). Esty, Goldstone, Gurr, Harff, Levy, Dabelko, Surko, and Unger (1999) report that wars and other violent political and ethnic action predict state failure. Bräutigam and Knack (2004) add that political instability and violence decreases the revenue available for public policy, due to a decrease in investment and trade as well as difficulty collecting taxes. Violence and political unrest also disrupt social, political, and economic priorities of both government and its citizens, which would affect the likelihood of educational policy adoption.

Finally, official development assistance (ODA) is an avenue by which substantial policy change could take place. Dollar and Easterly (1999) describe three mechanisms: first, aid could create opportunities for positive policy change through, for instance, capacity building; second, aid that is conditional on policy change could induce that change; third, aid that is conditional on promises of policy reform could induce reform. As government leaders around the globe have increasingly prioritized education, developmental aid might have contributed to the increase in CSLs in Africa.

From this prior literature, I identify a series of variables that can be important predictors of the adoption of compulsory schooling laws in Africa. I have organized the predictors to introduce them into the model by group, first testing classic diffusion and CSL diffusion predictors and then testing predictors that might be unique to the African context. From the policy diffusion literature, I include a time variable to track the likelihood of adoption over time and diffusion variables to capture the effects of the policy choices of nearby countries. The primary enrollment ratio, age-dependency ratio, and year of independence arise specifically in the CSL diffusion literature for Western nations. The adolescent fertility rate, GDP per capita, and international openness as measured by exports and imports as a percent of GDP come from the broader literature on mass education systems. Linguistic/historical ties, percent rural, population growth, total population size, levels of democracy, violence, death rate, and official development aid come from the literature more broadly on education in Africa.

## Data & Methods

The purpose of this study is to examine the predictors of lower secondary CSL adoption throughout Africa from 1960 to 2017. This year range is used because 1960 is one year prior to the first African country adopting such a law, and 2017 is the most recent available year for many of the variables. Data primarily come from the World Bank but also from other sources, such as UNESCO and the Demographic Health Surveys (DHS). For a full list of data sources, see Appendix A. Full descriptive statistics can be found in Appendix C.

### *Model*

I employ an event history analysis (EHA) model to examine the predictive factors for adoption of these lower secondary CSLs. A strength of EHA is that it combines internal determinants, such as economic characteristics that might predict policy adoption, with regional diffusion, which captures the influence of neighboring societies on the likelihood of adoption. A discrete time EHA examines a *risk set,* or the individuals/organizations/governments that are “at risk” of an event occurring, in this case, the adoption of lower secondary CSLs, in a given unit of time, which in this case is a year (Berry & Berry, 1990). The model is estimated on country-time data using a logistic[1] regression as follows:

$$\mathit{Threshold}_{ct} = \alpha + \beta_1\, \mathit{Year}_{ct} + \beta_2\, \boldsymbol{W}_{ct} + \beta_3\, \boldsymbol{X}_{ct} + \beta_4\, \boldsymbol{P}_{ct} + \beta_5\, \mathit{Diffusion}_{ct} + \varepsilon_{ct}$$

*Dependent Variable.* The dependent variable is *Threshold,* an indicator variable specifying whether the country adopted the CSL up to the given threshold in a given year. It is coded 1 in the year in which the country met the seven years of compulsory schooling threshold and 0 otherwise. According to the process of EHA, a country remains in the dataset, with an observation for every year that country "survives" (by not crossing the 7-year CSL threshold). The country drops out of the data set the year after it crosses the threshold. The full list of countries, CSL adoption years, and linguistic/historical ties is in Appendix B. Gabon was the first country to adopt CSLs that required at least 7 years of compulsory schooling, adopting in 1961, and Cote d'Ivoire was the last country to adopt, adopting in 2015. Fifteen countries had not adopted the CSL to the 7-year threshold by the end of 2017, which means that each of those countries that never adopted have observations for every year of the time frame. During data collection, I chose the year the CSL was adopted, instead of when the law went into effect, whenever possible. This means that the predictors themselves should only predict adoption of the law and should not themselves be affected by the CSL being in effect.

*Independent Variables. Year* is the year (from 1960 to 2017) of each observation of the country-year dataset. I tested multiple specifications of time, including linear, higher orders of the linear time variable as well as a series of year, half decade, and decade indicator variables. I use a linear specification of time due to issues with convergence of the models, perfect prediction of survival (which drops observations from the sample), and the fact that other models did not improve model fit enough to justify the more complex specifications.

*W* is a vector of linguistic/historical indicator variables (English, French, Arabic, and Portuguese), by country and year. I consider the linguistic tie to exist in a given country-year if the language is an official/governmental language and/or is commonly spoken. Very practically, this means that they share a common language, which means they can more easily communicate (directly and indirectly) about policy decisions and outcomes. It also captures cultural similarities, perhaps through common ancestors or cultural influences, that might shape their political philosophy and expectations surrounding children,

[1] All analyses are run in Stata14 (StataCorp, 2015). In additional models that are excluded from this paper, I use the Stata command `xtlogit` to test for unobserved heterogeneity. While the results do suggest there is unobserved heterogeneity in the adoption of these CSLs, the patterns that arise in the estimates are largely qualitatively similar, if not, more extreme. Accounting for the unobserved heterogeneity also removes some of the power, so I exclude these models from the results.

education, and government. The Arabic indicator is purely a linguistic variable, but the English, French, and Portuguese variables also capture historical ties. I consider the tie to exist historically for a given country if the country was colonized at any point by England, France, or Portugal, which were the three most common colonizers for these countries. This part of the linguistic/historical variables similarly captures how formal education systems were created in countries with common colonizers and how they relate to cultural and philosophical expectations about children, education, and government.

*X* is a vector of variables representing the demographic characteristics of primary enrollment ratio (ratio of total primary enrollment to population of the age group that corresponds to the primary grades), percent rural, population growth, age-dependency ratio (ratio of dependents, or people younger than 15 or older than 64, to the working-age population), adolescent fertility rate, GDP per capita, exports and imports as a percent of GDP, total population size, death rate, and foreign developmental assistance. I take the log of GDP per capita, exports and imports as a percent of GDP (trade), population size, and foreign developmental assistance (official developmental assistance, ODA) to increase normality of the variables.

*P* is a vector of political variables, including years since independence (or year of independence), political violence within the borders, and a continuous variable representing how democratic the country is in that year. The political violence measure is from the Center for Systemic Peace and is a rating from 0 (no episodes) to 10 (highest number of episodes) of international, civil, or ethnic violence within the country.

Diffusion captures the influence of other countries in the decision to adopt a policy. I am using two measures of diffusion: regional diffusion and linguistic/historical diffusion. Regional diffusion is the proportion of countries that share a border with the country of interest that adopted lower secondary CSLs by the year prior to the year of interest. Linguistic/historical diffusion is the proportion of countries that share a linguistic/historical tie with the country of interest that adopted lower secondary CSLs by the year prior to the year of interest.

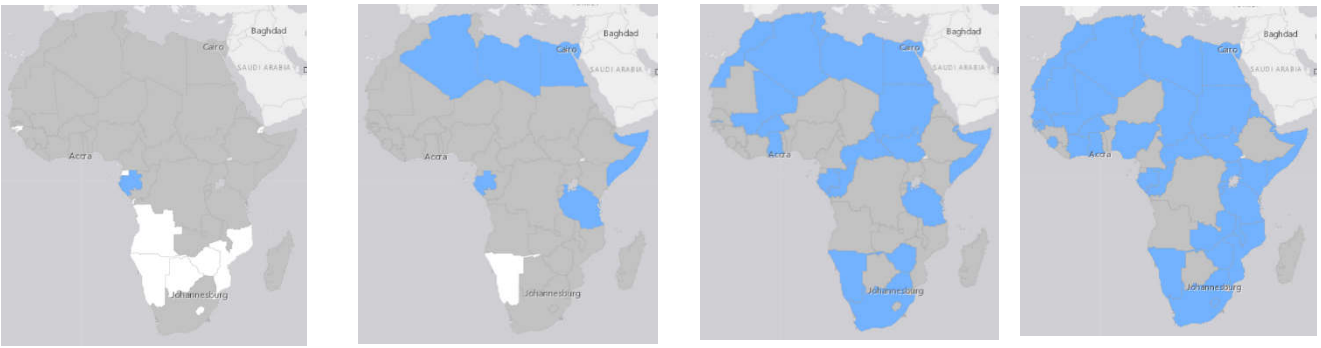
### *Data Availability*

Data availability is a major issue when conducting research in Africa. The World Bank and other sources have data on many, but not all, of the variables that are important to this analysis. Some potentially important information, like percent of immigrants, is not available for many of the country-years that are included in this analysis. Other variables that I would have liked to have include a demand for schooling, the demand for educated workers, prevalence of Christianity, ethnolinguistic fractionalization, and country resources spent on education. If these affect the probability of adoption, they are likely captured in some of the included variables, particularly time and linguistic/historical ties.

There is also missing data in several of the variables that are included. The variable with the highest missingness is the primary enrollment ratio because data collection did not start until the 1970s. It was an important variable identified from the literature, so I decided to include it in some models despite the high rate of missingness (30 percent). The variables imports and exports (15 percent), democracy (10 percent), and GDP per capita (9 percent) also have moderately high rates of missingness. All other variables are missing fewer than 5 percent of observations, often with missingness related to a single country. For instance, there is no primary enrollment data for Sudan in any year. I have obtained results that include and exclude each of these variables, and the resulting estimates remain relatively stable across models for each variable except linguistic/historical ties; across models, the estimates for linguistic/historical ties tend to maintain their direction but not magnitude.

## Results

I explore predictors of adoption of 7 years of CSLs in Africa between 1960 and 2017. Seven years encompasses all years of primary school and some or all years of lower secondary school for many countries, though it is only primary school for a few. Another way of doing this analysis would be to identify the exact number of years of schooling for lower secondary school in each country-year and base the analysis on that, but I chose to maintain constancy across the number of years instead of the level of school. I will still refer to these 7 years as encompassing lower secondary school throughout this paper. Figure 1 shows descriptively which countries had the CSLs throughout this time span. It shows that 39 out of 54 countries adopted lower secondary CSLs as of 2015.



1965 1985 2000 2015

Figure 1: Countries in Africa with CSLs

*Note: Countries in blue are those that have CSLs requiring at least seven years of compulsory schooling. Countries in gray are those that do not have CSLs requiring at least seven years of compulsory schooling. Countries in white are not yet independent.*

*Life Table Results*

Table 1 shows the adoptions in the form of a life table. The hazards are the probabilities that a country will adopt the CSLs in the specified time period, given that it did not experience it in any earlier time period (Singer & Willett, 2003). The hazards are fairly low in the 1900s and increase in likelihood over time, particularly as of 2004.

There have been many initiatives, conventions, and other activities over this period of time that could explain the patterns of adoption seen in Table 1. The Universal Declaration of Human Rights in 1948 and the related UNESCO Convention Against Discrimination in Education in 1960 both emphasized the importance of compulsory primary education, but their message did not spur much African CSL adoption that required at least 7 years. Neither did the 1966 International Covenant on Economic, Social and Cultural Rights, which reiterated that primary school should be compulsory, or the 1973 International Labour Organization Convention No. 138, which addressed the relationship between minimum working age and compulsory schooling.

These initiatives each built awareness for the topic of CSLs and paved the way for a plethora of activity in the 1980s, 1990s, and 2000s. The 1989 United Nations Convention on the Rights of the Child and connected 1990 African Charter on the Rights and Welfare of the Child appear to have prompted quite a few adoptions, and both reference compulsory primary or “basic” education, which often means primary but can also include lower secondary school. The adoptions that are likely related to these two events include Namibia, Ghana, and Tunisia in the early 1990s, based on when they signed or ratified the document and when they adopted the CSL (AU, 2017; RTE, 2014; UN, 1995).

Table 1: Life Table of Countries in Africa Adopting the CSLs, 1960-2017

| Year | Enter: *Countries that enter the risk set due to independence* | Risk Set: *Countries that have not yet adopted the CSLs in the given year but could* | Event: *Countries that adopted in the given year* | Hazard: *Proportion of countries that adopted in a year that could have adopted in that year* |
|---|---|---|---|---|
| 1960 | | 27 | | |
| 1961 | 1 | 28 | 1 | .04 |
| 1969 | 14 | 41 | 1 | .02 |
| 1975 | 6 | 46 | 1 | .02 |
| 1978 | 2 | 47 | 1 | .02 |
| 1979 | | 46 | 1 | .02 |
| 1981 | 1 | 46 | 2 | .04 |
| 1987 | | 44 | 1 | .02 |
| 1990 | 1 | 44 | 2 | .05 |
| 1992 | | 42 | 2 | .05 |
| 1993 | 1 | 41 | 1 | .02 |
| 1995 | | 40 | 1 | .03 |
| 1996 | | 39 | 2 | .05 |
| 1997 | | 37 | 2 | .05 |
| 2000 | | 35 | 2 | .06 |
| 2001 | | 33 | 2 | .06 |
| 2002 | | 31 | 1 | .03 |
| 2004 | | 30 | 4 | .13 |
| 2006 | | 26 | 1 | .04 |
| 2008 | | 25 | 2 | .08 |
| 2010 | | 23 | 3 | .13 |
| 2011 | 1 | 21 | 2 | .1 |
| 2012 | | 19 | 2 | .11 |
| 2013 | | 17 | 1 | .06 |
| 2015 | | 16 | 1 | .06 |
| 2017 | | 15 | 0 | |

Note: The CSLs of interest are those that require at least seven years of compulsory schooling.

The Education for All (EFA) initiative and the United Nations Millennium Declaration and accompanying Millennium Development Goals were exceedingly important. The World Declaration on Education for All adopted in Jomtien, Thailand, in 1990 and the Dakar Framework for Action from the World Education Forum in 2000 introduced and evaluated global and regional goals for the development of education. The Millennium Development Goals set educational goals and specific metrics with which to track them when they were declared in 2000. All of these goals brought immense awareness to education for development and accompanying resources to meet them, including the provision of compulsory primary education, which is explicitly stated in the Dakar Framework (UN, n.d.-a; UNESCO, 2000).

There are many additional initiatives that will likely maintain this momentum and encourage adoption moving forward. The Millennium Development Goals were retired in 2015 to make way for the Sustainable Development Goals, which also put weight on improved educational access and quality. Target 4.1, in particular, has an indicator that tracks whether all children are remaining in school until the end of lower secondary school, which will promote CSLs beyond primary school. These goals were adopted in 2015 and are intended to be completed by 2030 (UN, n.d.-b). For Africa specifically, there is the Continental Education Strategy for Africa 2016-2025 and the African Union's Agenda 2063 plan, which runs from 2013 to 2063.

*Event History Results*

Table 2 presents the results of the EHA analysis with odds ratios and statistical significance reported as p-values in parentheses underneath. Models 1-4 include classic regional diffusion, and Models 5-8 use linguistic/historic tie diffusion.[2] Models 1 and 5 capture basic policy diffusion models with time and diffusion. Models 2 and 6 include predictors that are prominent in the Western CSL diffusion and mass education literature. Models 3 and 7 look only at time, diffusion, and the variables identified from the education in Africa literature. Models 4 and 8 combine all variables into one model.

All of the geographical diffusion models present time as a statistically significant predictor of CSL adoption. The estimates are fairly stable across models and show that the odds of CSL adoption within a given year are 6-7 percent higher for each year that passes after 1960. Diffusion is either a nonsignificant predictor or negatively predicts adoption, suggesting that countries are less likely to adopt the CSLs of interest when more geographically proximal countries have already adopted them. Models 5-8 show that time is generally not a significant predictor of CSL adoption once the model controls for linguistic/historical diffusion. In fact, countries might be up to 1.2 times less likely to adopt as each year passes, according to Model 8. Linguistic/historical diffusion is important, unlike geographical diffusion, although the two diffusion variables are highly correlated at 0.64. Countries are 5 to 12 percent more likely to adopt the CSL for each percentage point increase in the percent of countries that share linguistic/historical ties that have already adopted the CSL.

[2] Coefficients for most variables are relatively stable across models that include various numbers and combinations of predictors.

Table 2: Event History Prediction, 1960-2017

| | Geographical Diffusion | | | | Linguistic/Historical Diffusion | | | |
|---|---|---|---|---|---|---|---|---|
| Model: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Year | 1.06<br>(0.001) | 1.06<br>(0.013) | 1.07<br>(0.002) | 1.06<br>(0.025) | 0.97<br>(0.265) | 0.97<br>(0.484) | 0.94<br>(0.188) | 0.86<br>(0.027) |
| Geographical Diffusion | 1.00<br>(0.978) | 1.00<br>(0.996) | 0.99<br>(0.102) | 0.98<br>(0.070) | | | | |
| Linguistic/ Historical Diffusion | | | | | 1.06<br>(0.001) | 1.05<br>(0.050) | 1.06<br>(0.007) | 1.12<br>(0.003) |
| Age-Dependency Ratio | | 1.01<br>(0.538) | | 1.00<br>(0.952) | | 1.02<br>(0.401) | | 1.02<br>(0.612) |
| Year of Independence | | 1.01<br>(0.454) | | 1.01<br>(0.526) | | 1.01<br>(0.402) | | 1.02<br>(0.273) |
| Adolescent Fertility Rate | | 0.99<br>(0.083) | | 0.99<br>(0.168) | | 0.99<br>(0.129) | | 0.99<br>(0.252) |
| Primary Enrollment Ratio | | 1.00<br>(0.715) | | 1.00<br>(0.995) | | 1.00<br>(0.666) | | 1.00<br>(0.650) |
| GDP per Capita (logged) | | 1.48<br>(0.161) | | 1.79<br>(0.168) | | 1.51<br>(0.150) | | 1.27<br>(0.590) |
| Trade as a % of GDP (logged) | | 0.60<br>(0.338) | | 0.78<br>(0.694) | | 0.64<br>(0.387) | | 0.73<br>(0.639) |
| English (0/1) | | | 0.51<br>(0.357) | 0.59<br>(0.593) | | | 0.53<br>(0.358) | 0.70<br>(0.700) |
| French (0/1) | | | 0.24<br>(0.044) | 0.28<br>(0.169) | | | 0.32<br>(0.092) | 0.49<br>(0.425) |
| Arabic (0/1) | | | 3.82<br>(0.007) | 3.71<br>(0.085) | | | 2.00<br>(0.188) | 1.10<br>(0.908) |
| Portuguese (0/1) | | | 0.27 | 0.52 | | | 1.63 | 11.31 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | (0.177) | (0.612) | | | (0.648) | (0.101) |
| Percent Rural | | | 0.98 | 1.01 | | | 0.98 | 0.99 |
| | | | (0.064) | (0.806) | | | (0.066) | (0.773) |
| Population Growth | | | 1.01 | 0.90 | | | 1.06 | 0.81 |
| | | | (0.930) | (0.749) | | | (0.709) | (0.567) |
| Population Size (logged) | | | 1.38 | 2.22 | | | 1.32 | 2.09 |
| | | | (0.064) | (0.004) | | | (0.096) | (0.008) |
| Death Rate | | | 0.95 | 0.97 | | | 0.97 | 0.96 |
| | | | (0.345) | (0.703) | | | (0.621) | (0.609) |
| Official Development Assistance (logged) | | | 1.00 | 1.75 | | | 1.12 | 1.58 |
| | | | (0.996) | (0.093) | | | (0.595) | (0.176) |
| Political Violence | | | 0.87 | 0.82 | | | 0.94 | 0.80 |
| | | | (0.288) | (0.343) | | | (0.632) | (0.277) |
| Democracy | | | 1.06 | 1.04 | | | 1.06 | 1.11 |
| | | | (0.361) | (0.679) | | | (0.404) | (0.290) |
| Observations | 2,043 | 1,219 | 1,800 | 1,112 | 2,043 | 1,219 | 1,800 | 1,112 |
| Pseudo $R^2$ | 0.06 | 0.10 | 0.13 | 0.18 | 0.09 | 0.12 | 0.15 | 0.20 |
| Log Likelihood | -180.84 | -123.07 | -159.63 | -107.08 | -174.90 | -120.74 | -157.46 | -104.22 |

Exponentiated coefficients; P-values in parentheses

From the Western CSL diffusion literature, the age-dependency ratio, year of independence, and primary enrollment ratio do not predict adoptions, which diverge from Richardson (1980), Meyer, Ramirez, and Soysal (1992), and Fyfe (2005), respectively. The adolescent fertility rate negatively predicts adoption with a marginally significant estimate in Model 2, which means that countries are less likely to adopt if their females aged 15 to 18 are having more children. GDP per capita and trade are not significant predictors of adoption, but their directions suggest that a higher GDP per capita positively predicts adoption, likely due to lesser budget constraints (Murtin & Viarengo, 2011), and trade negatively predicts adoption. The trade finding is in direct contrast to Murtin and Viarengo (2011), who find that trade openness predicts more years of CSLs in Western Europe.

The results for the linguistic/historical ties suggest that these groupings are important in predicting adoption. In particular, the results suggest that countries with French and perhaps English and Portuguese ties are up to 4.2 times less likely to adopt the CSL than countries without those ties. Countries with Arabic ties are much more likely to adopt the CSL than countries without Arabic ties. In Model 3, countries with Arabic ties are almost 3 times more likely to adopt the CSL than countries without the Arabic ties. The estimates for Portuguese ties vary substantially across models because of the small proportion of countries with these ties. I will note that the only country-years that do not have English, French, Arabic, or Portuguese ties are Equatorial Guinea before 1997 and Ethiopia before 1994.

The other variables identified in the literature on education in Africa are, for the most part, not significant predictors of the CSL adoption. The biggest exception is the population size, where the estimates suggest that adoption is more likely as the population size increases. In Models 3 and 7, adoption is less likely in countries with a higher percent rural, and in Model 4, official development assistance positively predicts adoption. While they are not significant, the direction of the political violence estimates suggest that adoption might be less likely in countries with more political violence.

## Discussion

This study uses an event history analysis to examine the predictors of the adoption of a law requiring 7 years of compulsory schooling throughout Africa. The results suggest that time, linguistic/historical ties and their diffusion, and population size are the most important predictors. Of the other predictors, some – such as the adolescent fertility rate – predict CSL adoption in line with what would be expected from prior Western CSL diffusion literature, while others – such as trade openness – conflict with prior research. Several variables obtained from the African education and policymaking literature also proved to be valuable, adding that adoption is more likely with a larger population size and more ODA and is potentially less likely with more political violence.

The traditional diffusion models show that as time passes, the countries that had not already adopted the CSLs were increasingly likely to adopt them. This reflects the situation in the 1960s and 1970s, with many African countries becoming independent and building up their governmental and educational systems that needed time to mature before adopting CSLs of so many years. Over time, communication has become easier, which has made governance problems and subsequent international pressures more public. International attention has also been particularly focused on access to education since 1989 and has only increased further since then.

The relationship between linguistic/historical ties and adoption is likely more complex. Countries with Arabic ties were more likely to adopt than those without, which means they were more likely to adopt the CSL closer to their year of independence than were other countries. Countries with any of the included linguistic/historical ties were also more likely to adopt when a higher percent of countries in their same group had already adopted. This could be due to the classic diffusion theory explanation, which is that countries with the same linguistic/historical ties communicated frequently enough to make decisions at similar times or observe each other's decisions and follow their lead. However, the bond between people in different countries who share linguistic/historical ties would have to be incredibly strong for the decades in which communication was not as instantaneous in Africa for there to be a significant effect of linguistic/historical diffusion but not regional diffusion. More likely, similarities in the existing governmental and educational systems from a relatively common language/history are why the results show such an effect for linguistic/historical ties and diffusion. This is just speculation, however, that should be explored in future research.

A surprising finding is that the other variables were nonsignificant in most of the models. In particular, it is surprising that democracy and political violence do not significantly predict the adoption of these CSLs. Perhaps this alludes to the disconnect between policy adoption and practice. While set in the theory of action for CSLs, this analysis does not tell us anything about actual enrollment or attendance rates in these compulsory grades. For instance, parents might be hesitant to send their children to compulsory school if there are low private returns to schooling, which would be particularly true if the family needs another source of income from the child in the present more than they need that child's future, potentially higher due to higher levels of education, income (Akkari, 2004; Lloyd & Blanc, 1996).

There are a variety of limitations to consider in this paper, primarily around data availability. I was not able to find information for every variable that might have been an important predictor of the adoption of these CSLs, like percent of immigrants in the population, as per the prior CSL diffusion literature. There are high rates of missingness in several variables, including primary enrollment ratio, that reduce the sample size for the models in which they are included. Additionally, there could be errors in the data available through my own data collection of the years of CSL adoption and the data available from the World Bank and other sources. The lack of or errors in potentially important variables and high rates of missingness for some variables are clear limitations of my study. Data availability and quality are often major roadblocks in analyses that are historical and/or focus on Africa. The variables included present a large proportion of those that would be used in an ideal analysis, and most estimates remain stable across models with different covariates, so my analysis is optimal for this point in time, despite its flaws.

## Conclusion

This study seeks to identify important predictors in the diffusion of lower secondary CSLs throughout Africa. The countries have been adopting these CSLs since the 1960s, and there is still space for more to do so today. The CSL diffusion literature is sizeable but fails to fully consider contexts outside of Western Europe and the United States. This paper helps to fill that gap and provides insight into the policy adoption process on the continent of Africa.

Researchers and policymakers can use the results from this study to gain insight on the policymaking process in Africa and how it might differ from the process in areas of the world that are currently more thoroughly researched. In particular, my study suggests that, at least before widespread governmental access to communications technology, part of the policymaking process in Africa was/is driven by common systems that mature to be ready to accept new policies. This means that those who want to adopt new policies should attend to the systems and structures already in place to ensure that they are ready to accept the new policy. This study does not provide insight on what exactly that means except to say that these CSLs were adopted following patterns primarily based on commonalities in language and/or history. Additional research can probe into these two paths further, discovering how communication allowed due to common languages and/or similarities in governance systems contribute to the policy learning and decision-making of governments.

Future research should also focus on the enforcement of CSLs and their quality. Exceptions to the CSLs and leniency (including lack of proper enforcement) weaken the effectiveness of these laws in improving school attendance (Oreopoulos, 2009). For instance, a nationally representative survey of South African youth revealed that 23 percent of 16-24 year-olds had not successfully completed their compulsory education (Operario, Cluver, Rees, MacPhail, & Pettifor, 2008). CSLs can therefore be seen as a way to encourage parents to voluntarily send their children to school and to provide the government an opportunity to enforce compliance with compulsory education rather than as a guarantee of increased attainment. As for quality, there can be a very real tradeoff between providing broad access to education and providing high-quality education (Deininger, 2003). A low-quality education might provide a custodial function, which could be useful to keep children out of risky situations and provide childcare for families, but a high-quality education is necessary to maximize the use and usefulness of schooling. The quality of schooling also drives demand, leading Michaelowa (2001) to proclaim that, “Ensuring education quality is a necessary complement to enrolment: quality and quantity have to go hand in hand” (pg. 2). Successful CSLs, therefore, require incentivizing families to send their children to school with the work of enforcement officers and a guarantee that school will benefit the child and the child's family (Fyfe, 2005; Landes & Solmon, 1972; Murtin & Viarengo, 2011).

Compulsory schooling laws are seen on the international stage as a positive step a government can take to protect and enrich the lives of a country's population, and official international development assistance does seem to predict CSL adoption. However, a few of the nonsignificant results combined with prior literature tells a story about how there might be a disconnect between policy adoption and practice in some governments. Without the capacity and/or will to uphold CSLs, they are simply a symbolic gesture that, cynically, could only signal to development organizations that the country is worthy of continued support. On a more positive note, that symbolic gesture can promote a national priority to support education for all children and improve access and attendance.

## REFERENCES

Acemoglu, D., & Angrist, J. (2000). How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws. *NBER Macroeconomics Annual*, *15*, 9–59. https://doi.org/10.1086/654403

ADB. (2002). *Achieving the Millennium Development Goals in Africa: Progress, Prospects, and Policy Implications*.

Akkari, A. (2004). Education in the Middle East and North Africa: The current situation and future challenges. *International Education Journal*, *5*(2), 144–153.

Astiz, M. F., Wiseman, A. W., & Baker, D. P. (2002). Slouching towards Decentralization: Consequences of Globalization for Curricular Control in National Education Systems. *Comparative Education Review*, *46*(1), 66–88. https://doi.org/10.1086/324050

AU. (2017). *List of Countries Which Have Signed, Ratified/Acceded to the African Youth Charter*. Addis Ababa.

Bandiera, O., Mohnen, M., & Rasul, I. (2015). *Nation-Building Through Compulsory Schooling During the Age of Mass Migration*. *Unpublished*. Retrieved from http://sticerd.lse.ac.uk/dps/eopp/eopp57.pdf%5Cnpapers3://publication/uuid/244D4CEC-48DE-4251-BDD4-53EE10FD0348

Bennell, P. (1996). Rates of return to education: Does the conventional pattern prevail in sub-Saharan Africa. *World Development*, *24*(1), 183–199.

Berry, F. S., & Berry, W. D. (1990). State Lottery Adoptions as Policy Innovation: An Event History Analysis. *American Political Science Review*, *84*(2), 395–415.

Black, S. E., Devereux, P., & Salvanes, K. G. (2008). Staying in the Classroom and out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births. *The Economic Journal*, *118*(530), 1025–1054.

Bräutigam, D. A., & Knack, S. (2004). Foreign Aid, Institutions, and Governance in Sub-Saharan Africa, *52*(2), 255–285.

Brunello, G., Fort, M., & Weber, G. (2009). Changes in Compulsory Schooling, Education and the Distribution of Wages in Europe. *The Economic Journal*, *119*(536), 516–539.

Carr-Hill, R., Eshete, A., Sedel, C., & de Souza, A. (2005). *The education of nomadic people in East Africa: Djibouti, Eritrea, Ethiopia, Kenya, Tanzania and Uganda*. Retrieved from http://eprints.whiterose.ac.uk/72739/

Deininger, K. (2003). Does cost of schooling affect enrollment by the poor? Universal primary education in Uganda. *Economics of Education Review*, *22*(3), 291–305. https://doi.org/10.1016/S0272-7757(02)00053-5

Dollar, D., & Easterly, W. (1999). The search for the key: Aid, investment and policies in Africa. *Journal of African Economies*, *8*(4), 546–577. https://doi.org/10.1093/jae/8.4.546

Eckstein, Z. (1994). The effects of compulsory schooling on growth, income distribution and welfare. *Journal of Public Economics*, *54*, 339–359.

Edwards, L. N. (1978). An Empirical Analysis of Compulsory Schooling Legislation, 1940-1960. *The Journal of Law & Economics*, *21*(1), 203–222.

Elbadawi, I., & Sambanis, N. (2000). Why Are There So Many Civil Wars in Africa? Understanding and Preventing Violent Conflict. *Journal of African Economies*.

Esty, D. C., Goldstone, J. A., Gurr, T. R., Harff, B., Levy, M., Dabelko, G. D., … Unger, A. N. (1999). State Failure Task Force Report: Phase II Findings. *Environmental Change & Security Project Report*, (5).

Fataar, A. (1997). Access to Schooling in a Post-Apartheid South Africa: Linking Concepts to Context. *International Review of Education*, *43*(4), 331–348.

Fuller, B., & Rubinson, R. (1992). *The Political Construction of Education: The State, School Expansion,* *and Economic Change.* New York: Praeger Publishers.

Fyfe, A. (2005). *Compulsory Education and Child Labour: Historical Lessons, Contemporary Challenges and Future Directions*. Retrieved from http://www.ilo.org/ipecinfo/product/viewProduct.do;jsessionid=620d8a04f9f69767dd24d95e756ec452bc8bdb96af85d594080b1a3ef72a6062.e3aTbhuLbNmSe3qNay0?productId=1099

Hill, L. D., Baxen, J., Craig, A. T., Namakula, H., & Sayed, Y. (2012). Citizenship, social justice, and evolving conceptions of access to education in South Africa: Implications for research. *Review of Research in Education*, *36*(1), 239–260. https://doi.org/10.3102/0091732X11421461

Hill, M. A., & King, E. M. (1993). *Women's education in developing countries: An overview*.

Landes, W. M., & Solmon, L. C. (1972). Economic History Association Compulsory Schooling Legislation: An Economic Analysis of Law and Social Change in the Nineteenth Century. *The Journal of Economic History*, *32*(1), 54–91.

Lang, K., & Kropp, D. (1986). Human Capital Versus Sorting: The Effects of Compulsory Attendance Laws. *The Quarterly Journal of Economics*, *101*(3), 609–624. https://doi.org/10.1080/02724980343000242

Lewin, K., & Sabates, R. (2011). *Changing patterns of access to education in Anglophone and Francophone countries in sub Saharan Africa: Is education for all pro-poor?* Retrieved from http://sro.sussex.ac.uk/30025/

Lloyd, C B, Kaufman, C. E., & Hewett, P. (2000). The spread of primary schooling in Sub-Saharan Africa: implications for fertility change. *Population and Development Review*, *26*(3), 483–515. https://doi.org/10.1111/j.1728-4457.2000.00483.x

Lloyd, Cynthia B, & Blanc, A. K. (1996). Children's Schooling in sub-Saharan Africa: The Role of Fathers, Mothers, and Others. *Population and Development Review*, *22*(2), 265–298.

Meyer, J. W., Ramirez, F., & Soysal, Y. N. (1992). World Expansion of Mass Education, 1870-1980. *Sociology of Education*, *65*(2), 128–149.

Michaelowa, K. (2001). Primary Education Quality in Francophone Sub-Saharan Africa: Determinants of Learning Achievement and Efficiency Considerations. *World Development*, *29*(10).

Murtin, F., & Viarengo, M. (2011). The Expansion and Convergence of Compulsory Schooling in Western Europe, 1950-2000. *Economica*, *78*(311), 501–522. https://doi.org/10.1111/j.1468-0335.2009.00840.x

Nawaz, M., & Tanveer, S. (1975). Compulsory Education: National and International Perspective. *Educational Leadership*, *32*(4), 278–282. Retrieved from http://www.ascd.com/ASCD/pdf/journals/ed_lead/el_197501_nawaz.pdf

NECC. (1989). *Report of the national education conference: Consolidate and advance to People's Education*. Cape Town.

Operario, D., Cluver, L., Rees, H., MacPhail, C., & Pettifor, A. (2008). Orphanhood and completion of compulsory school education among young people in South Africa: Findings from a national representative survey. *Journal of Research on Adolescence*, *18*(1), 173–186. https://doi.org/10.1111/j.1532-7795.2008.00555.x

Oreopoulos, P. (2006a). Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter. *The American Economic Review*, *96*(1), 152–175.

Oreopoulos, P. (2006b). The compelling effects of compulsory schooling: Evidence from Canada. *Canadian Journal of Economics*, *39*(1), 22–52. https://doi.org/10.1111/j.0008-4085.2006.00337.x

Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics*, *91*(11–12), 2213–2229. https://doi.org/10.1016/j.jpubeco.2007.02.002

Oreopoulos, P. (2009). Would more compulsory schooling help disadvantaged youth? Evidence from recent changes to school-leaving laws. In J. Gruber (Ed.), *The Problems of Disadvantaged Youth: An*

*Economic Perspective* (pp. 85–112). University of Chicago Press. Retrieved from http://www.nber.org/chapters/c0588

Oreopoulos, P., Page, M. E., & Stevens, A. H. (2006). The intergenerational effects of compulsory schooling. *Journal of Labor Economics*, *24*(4), 729–760. https://doi.org/10.1086/506484

Pischke, J., & von Wachter, T. (2008). Zero returns to compulsory schooling in Germany: Evidence and interpretation. *The Review of Economics and Statistics*, *90*(3), 592–598. https://doi.org/10.1162/rest.90.3.592

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. *Education Economics*, *26*(5), 445–458. https://doi.org/10.1080/09645292.2018.1484426

Ramirez, F. O. (1989). Reconstituting Children: Extension of Personhood and Citizenship. In D. Kertzer & K. W. Schaie (Eds.), *Age Structuring in Comparative Perspective* (pp. 143–165). Hillsdale.

Ramirez, F. O., & Boli, J. (1987). The Political Construction of Mass Schooling: European Origins and Worldwide Institutionalization. *Sociology of Education*, *60*(1), 2–17.

Richardson, J. G. (1980). Variation in Date of Enactment of Compulsory School Attendance Laws: An Empirical Inquiry. *Sociology of Education*, *53*(3), 153–163.

RTE. (2014). *International Instruments*. Retrieved from http://www.right-to-education.org/sites/right-to-education.org/files/resource-attachments/RTE_International_Instruments_Right_to_Education_2014.pdf

Schultz, T. P. (1999). Health and Schooling Investments in Africa. *The Journal of Economic Perspectives*, *13*(3), 67–88.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Stasavage, D. (2005). Democracy and education spending in Africa. *American Journal of Political Science*, *49*(2), 343–358.

StataCorp. (2015). Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Tikly, L. (2001). Globalisation and Education in the Postcolonial World: towards a conceptual framework. *Comparative Education*, *37*(2), 151–171. https://doi.org/10.1080/03050060120043394

Tyack, D. B. (1976). Ways of Seeing: An Essay on the History of Compulsory Schooling. *Harvard Educational Review*, *46*(3), 355–390.

UN. (n.d.-a). Millennium Development Goals Background. Retrieved from https://www.un.org/millenniumgoals/bkgd.shtml

UN. (n.d.-b). Sustainable Development Goals. Retrieved from https://sustainabledevelopment.un.org/?menu=1300

UN. (1995). *Convention on the Rights of the Child*. New York.

UN. (2015). *The Millennium Development Goals Report*. https://doi.org/978-92-1-101320-7

UNESCO. (1961). *Conference of African States on the Development of Education in Africa: Final Report*. https://doi.org/http://www.royalcommission.vic.gov.au/finaldocuments/summary/PF/VBRC_Summary_PF.pdf

UNESCO. (2000). *The Dakar Framework for Action*. Paris. Retrieved from http://unesdoc.unesco.org/images/0012/001211/121147e.pdf

Van Horn Melton, J. (1988). Introduction. In *Absolutism and eighteenth-century origins of compulsory schooling in Prussia and Austria*. Cambridge: Cambridge University Press.

WCEFA. (1990). *Meeting Basic Learning Needs: A Vision for the 1990s*. *World Conference on Education for All*. New York.

Weiner, M. (1991). *The child and the state in India: Child labor and education policy in comparative perspective*. Princeton University Press.

## APPENDIX A: LIST OF DATA SOURCES

| **Variable** | **Data Source** |
|---|---|
| Years of compulsory education | Variety of sources, including World Bank, TIMSS encyclopedia, Library of Congress country studies, The Guide to Higher Education in Africa (2004), documents from the US Departments of State and Labor, other governmental documents, and personal communications with African nationals. |
| Linguistic/historical ties and year of independence | Variety of sources, including Lewin & Sabetes (2011), World Bank, CIA Factbook, www.nationsonline.org, and www.answersafrica.com. |
| Demographic variables | World Bank |
| Political violence & Democracy | Center for Systemic Peace |

## APPENDIX B: LIST OF CSL ADOPTION YEARS AND LINGUISTIC/HISTORICAL TIES

| Country | Year Adopted | Linguistic/Historical Ties |
|---|---|---|
| Algeria | 1975 | Arabic & French |
| Angola | Not adopted by 2017 | Portuguese |
| Benin | Not adopted by 2017 | French |
| Botswana | Not adopted by 2017 | English |
| Burkina Faso | 1996 | French |
| Burundi | 2012 | French, English & French as of 2014 |
| Cameroon | Not adopted by 2017 | English & French |
| Cape/Cabo Verde | 2010 | Portuguese |
| Central African Republic | 1997 | French |
| Chad | 2006 | Arabic & French |
| Comoros | Not adopted by 2017 | Arabic & French |
| Congo | 1992 | French |
| Cote d'Ivoire | 2015 | French |
| Democratic Republic of the Congo | Not adopted by 2017 | French |
| Djibouti | 2000 | Arabic & French |
| Egypt | 1981 | Arabic & English |
| Equatorial Guinea | Not adopted by 2017 | None, French as of 1997, French & Portuguese as of 2007 |
| Eritrea | 2010 | Arabic & English |
| Ethiopia | Not adopted by 2017 | None, English as of 1994 |
| Gabon | 1961 | French |
| The Gambia | 1997 | English |
| Ghana | 1992 | English |
| Guinea | Not adopted by 2017 | French |
| Guinea-Bissau | 2001 | Portuguese |
| Kenya | 2002 | English |
| Lesotho | 2010 | English |
| Liberia | Not adopted by 2017 | English |
| Libya | 1969 | Arabic |
| Madagascar | Not adopted by 2017 | French, English & French 2007 to 2010 |
| Malawi | 2013 | English |
| Mali | 1993 | French |
| Mauritania | 2001 | Arabic & French |
| Mauritius | 2004 | English & French |
| Morocco | 2000 | Arabic & French |
| Mozambique | 2011 | Portuguese |
| Namibia | 1990 | English |

| | | |
|---|---|---|
| Niger | Not adopted by 2017 | French |
| Nigeria | 2004 | English |
| Rwanda | 2008 | French, English & French as of 1995 |
| Sao Tome and Principe | Not adopted by 2017 | Portuguese |
| Senegal | 2004 | French |
| Seychelles | 1981 | English & French |
| Sierra Leone | 2004 | English |
| Somalia | 1979 | Arabic |
| South Africa | 1996 | English |
| South Sudan | 2012 | English |
| Sudan | 1995 | Arabic & English |
| Swaziland | Not adopted by 2017 | English |
| Tanzania | 1978 | English |
| Togo | Not adopted by 2017 | French |
| Tunisia | 1990 | Arabic & French |
| Uganda | 2008 | English |
| Zambia | 2011 | English |
| Zimbabwe | 1987 | English |

## APPENDIX C: SAMPLE MEANS AND STANDARD DEVIATIONS ACROSS FULL SAMPLE, YEARS POOLED

|  | Mean | Standard Deviation | Minimum | Maximum | Number of Observations |
|---|---|---|---|---|---|
| Regional Diffusion | 23% | 28% | 0% | 100% | 2,043 |
| Linguistic/Historical Diffusion | 19% | 20% | 0% | 81% | 2,043 |
| English Ties | 40% | 49% | 0% | 100% | 2,043 |
| French Ties | 52% | 50% | 0% | 100% | 2,043 |
| Arabic Ties | 17% | 38% | 0% | 100% | 2,043 |
| Portuguese Ties | 10% | 30% | 0% | 100% | 2,043 |
| Primary Enrollment Ratio | 79.61 | 32.12 | 11.71 | 173.82 | 1,427 |
| Percent Rural | 72% | 15% | 23% | 98% | 2,043 |
| Population Growth | 2.64 | 1.16 | -6.18 | 11.81 | 2,043 |
| Age Dependency Ratio | 91.31 | 10.39 | 46.12 | 114.33 | 2,043 |
| Adolescent Fertility Rate (per 1,000 women ages 15-19) | 137.32 | 49.86 | 19.72 | 232.48 | 2,043 |
| GDP per Capita (current US$) | $738.00 | $1,614.52 | $37.52 | $22,742.38 | 1,828 |
| *GDP per Capita (logged)* | *5.96* | *0.98* | *3.62* | *10.03* | *1,828* |
| Trade as a % of GDP | 67% | 45% | 6% | 532% | 1,739 |
| *Trade (logged)* | *4.05* | *0.54* | *1.84* | *6.28* | *1,739* |
| Total Population Size | 11,262,356 | 17,067,327 | 60,504 | 135,393,616 | 2,043 |
| *Population Size (logged)* | *15.38* | *1.44* | *11.01* | *18.72* | *2,043* |
| Year of Independence | 1959 | 22 | 1847 | 2011 | 1,985 |
| Political Violence | 0.79 | 1.74 | 0.00 | 10.00 | 1,944 |
| Official Development Assistance (per capita) | 44.71 | 62.64 | -11.97 | 664.98 | 2,005 |
| *ODA (logged)* | *3.06* | *1.43* | *-4.80* | *6.50* | *2,004* |
| Death Rate | 16.29 | 5.89 | 5.45 | 38.49 | 2,043 |
| Democracy | 2.03 | 3.07 | 0.00 | 10.00 | 1,838 |

Note: These descriptive statistics are calculated from only the country-years in the estimation sample, which excludes all country-years after a particular country adopted CSLs that have at least 7 years of compulsory schooling.

# CHAPTER 2

## TEACHING WITHOUT BOUNDARIES:
## ADAPTING COLLABORATIVE INQUIRY CYCLES TO THE AMERICAN CONTEXT

The past few decades revealed renewed interest in teacher collaboration to support professional development and working conditions as well as improve student learning (Achinstein, 2002; Bond, 2014; Bruce, Flynn, & Bennett, 2016; Goddard, Goddard, & Tschannen-Moran, 2007). However, just as many education reforms fade after initial implementation, often teacher collaboration initiatives fail to take root in the day-to-day operation of schools after implementation teams remove their supports (Giles & Hargreaves, 2006; Sindelar, Shearer, Yendol-Hoppey, & Liebert, 2006; Zech, Gause-Vega, Bray, Secules, & Goldman, 2000). This phenomenon has prompted questions about the decision-making processes among key stakeholders regarding local adoption and adaptation that are necessary for educational reforms to be successful over time.

As in many parts of the United States, teachers in Tennessee schools have been encouraged to work collaboratively and build communities of practice (Papay, Taylor, Tyler, & Laski, 2016). In this study, I focus on a teacher collaboration model called Teacher Peer Excellence Groups (TPEGs; Cravens, Drake, Goldring, & Schuermann, 2017), which was designed to capture components of the Japanese lesson study and Chinese teaching-study groups (Fujii, 2016; Huang & Shimizu, 2016; Jensen, Sonnemann, Roberts-Hull, & Hunter, 2016; Lewis, Perry, & Murata, 2006) and adapt them to the very different American educational context. For the TPEG model, groups of teachers meet routinely for a disciplined cycle of collaborative lesson planning, observe each other deliver the lesson, provide feedback and updates to the lesson, and store the lesson to share with other teachers and access in future years. Since its pilot implementation in 2013, some schools and/or teachers have adapted this TPEG process to better fit their needs and now have a collaborative planning system that is fully integrated into their regular practice.

This study examines this support and adaptation process in depth to explore what happens in practice when a collaborative inquiry model is implemented in schools. I conduct a series of exploratory case studies to examine the supports and processes related to the TPEG implementation and adaptation in three pilot schools in Tennessee. These three schools were purposively selected from 27 pilot schools because they have sustained parts of the TPEG model to varying extents. The research questions for this study are: Why have some schools or some teams of teachers continued to use collaborative inquiry cycles while others have stopped? Compared to its original theory of action, how has TPEG evolved to the local context? The first question addresses the enabling conditions for collaborative inquiry cycles, and the second examines the adaptation of TPEG into collaborative practices that were sustained longer than TPEG in these three schools. This research can shed light on the appropriateness of collaborative inquiry cycles as a key feature for teacher communities of practice and ways to make them successful and sustainable in public schools.

## Collaborative Inquiry Cycles

Collaborative inquiry cycles as a mechanism to improve teaching and learning have been in practice in Asia for decades. For example, Japanese lesson study is over a century old and includes five steps: goal setting, lesson planning, research lesson, post-lesson discussion, and reflection (Fujii, 2016). Chinese lesson study has been practiced across the country since the 1950s, and its core components are studying lessons, development, and public lessons (Huang & Shimizu, 2016). The point of these systems is to collaborate and make the lesson planning process visible so that teachers can build on each other's knowledge and prior work (Fujii, 2016; Huang & Shimizu, 2016). These cycles allow for teachers to improve both their practice and student learning (OECD, 2011).

### *Theory of Action*

Successful collaborative relationships in education are often spontaneous, voluntary, and unpredictable (Hargreaves, 1994), which is in direct opposition to deliberate inquiry cycles. Collaborative inquiry cycles involve teams of teachers organized by subject matter and/or grade level, deeply engaged in collaborative lesson planning, giving and receiving actionable feedback about the lessons, and holding one another accountable for improving their practice. The theory of change for the collaborative inquiry model demonstrates the movement from practitioner sharing to professional knowledge base building (Hiebert, Gallimore, & Stigler, 2002; Stigler & Hiebert, 2016) through three non-negotiable objectives. First is public: teachers make their practice public through collaborative lesson planning, peer observations, and peer feedback. Second is that the materials and expertise gathered during an inquiry cycle are cumulative, accessible, and shareable to other teachers so that teachers do not have to reinvent the wheel every time they get a new teaching assignment. Third, there is a mechanism for validation and improvement of the lesson from experts, which happens during the peer feedback sessions. Figure 1 shows the conceptual relationship between the steps of collaborative inquiry cycles (lesson planning, observation, feedback and refinement, and lesson storage and sharing), enabling conditions, and instructional improvement.

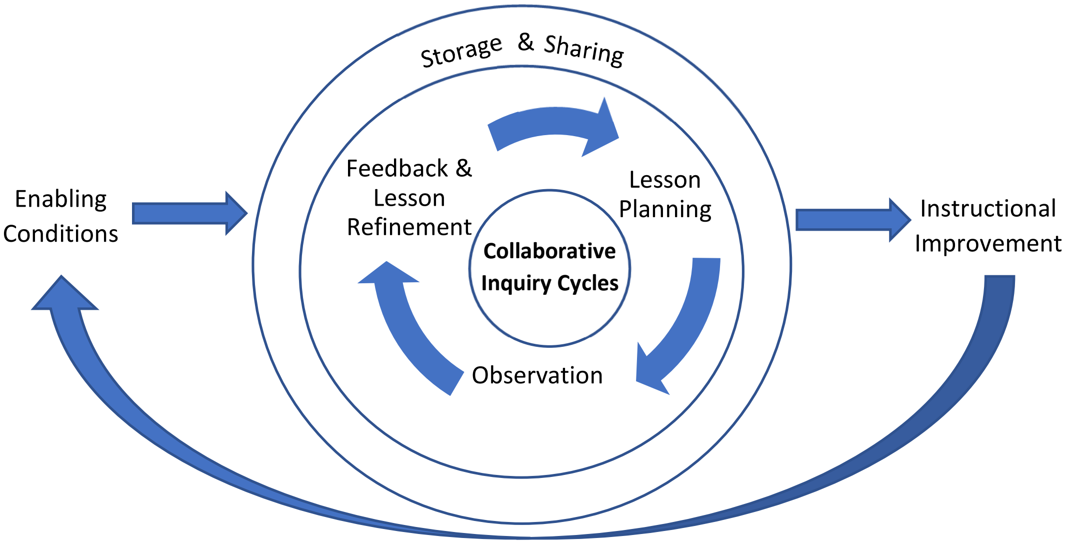

Storage & Sharing
Enabling Conditions
Feedback & Lesson Refinement
Lesson Planning
Collaborative Inquiry Cycles
Observation
Instructional Improvement

**Figure 1: Conceptual Relationship Between the Aspects of Collaborative Inquiry Cycles**
Adapted from Cravens & Drake, 2017, and Cravens et al., 2017

The theory of action provides that the steps of the collaborative inquiry cycle will lead to instructional collaboration and deprivatized practice, which will result in instructional improvement, as suggested by Hiebert, Gallimore, and Stigler (2002). They write that researchers' general but generalizable knowledge and teachers' "craft" and contextual knowledge are often incompatible but that these forms of knowledge can be integrated to move educational practice from practitioner knowledge to professional knowledge. Hiebert, Gallimore, and Stigler (2002) assert that if practitioner knowledge can be public, verified, improved, stored, and shared, it can be transformed into a professional knowledge base that improves the practice of teaching.

### *Enabling Conditions of Collaborative Inquiry Cycles*

Other education systems have taken note of the success of the lesson study model and have used lesson-drawing (Rose, 1991) to take certain characteristics of the model and adapt them to new environments. That process is neither easy nor straightforward, as there is no guarantee that the lessons that are drawn are going to be sensible, necessary, or sufficient to allow for the collaboration to be successful. According to the literature and the theory of action, instructional leadership, professional community, trust, and efficacy are all important when attempting to implement collaborative inquiry cycles. During data analysis, I examine each of these enabling conditions in the context of the three study schools.

School leadership is often thought to be the second most important school factor that affects student learning, only after classroom instruction. Leaders are key stakeholders who affect student learning by making numerous daily decisions. They influence the school organization and the people within it, including (re)designing the structure of the school, cultivating expectations and school culture, developing teachers' pedagogical and content knowledge, and cultivating professional communities (Coburn, 2001; Goldring, Porter, Murphy, Elliott, & Cravens, 2009; Leithwood, Seashore, Anderson, & Wahlstrom, 2004; Printy, 2008; ten Bruggencate, Luyten, Scheerens, & Sleegers, 2012; Youngs & King, 2002). Principals that focus specifically on improving classroom practices are considered 'instructional leaders' (Hallinger & Heck, 2010). Using a meta-analysis, Robinson, Lloyd, & Rowe (2008) find that instructional leadership has a stronger influence on student outcomes than do other types of leadership, such as transformational leadership, for which leaders focus on inspiring staff to better engage with their work. Instructional leaders can therefore use their influence to shape the structural factors, such as time, that are necessary to support collaborative inquiry cycles and direct teacher efforts toward student learning.

Leadership supports and affects many parts of the school environment, including its professional community, the second enabling condition. Broadly, a professional community consists of teachers who frequently interact using a set of shared norms about improving teaching and learning. More specifically, teachers in a professional community reflect on instructional practices and student learning, observe each other's teaching practices, problem solve together, and share work through peer collaboration (Bryk, Camburn, & Louis, 1999). In a study of 24 schools, Louis, Marks, and Kruse (1996) find that professional communities have a positive relationship to teachers taking responsibility for student learning, and Louis and Marks (1998) report that professional communities positively affect classroom organization and student academic performance.

School culture and climate, including trust and respect among the faculty, are important in creating a professional community (Bryk et al., 1999; Louis et al., 1996). This leads to a third condition, trust within the school, primarily between teachers on the same collaborative inquiry teams. Bryk & Schneider (2002) and Louis (2006) argue that trust within schools is essential to facilitate daily practice within a school and improvement measures. They write that trust between faculty members in a school is built on respect, competence, a personal regard for other people, integrity, and agreement on issues such as what students should learn and how teachers should instruct and behave. Effective principals can improve their schools by building trust between their teachers (Youngs & King, 2002), or teachers can develop trust themselves over time in a way that allows them to work together well and take full advantage of the benefits that can come from collaborative inquiry cycles.

The last enabling condition I will cover is efficacy, which is the teacher's self-belief that s/he can focus his/her efforts to successfully attain an educational goal. Pedagogical efficacy, for instance, means that teachers feel like they can successfully integrate new practices, like collaborative inquiry cycles, into their regular practice (Bandura, 1997). Bandura (1997) asserts that teachers use four types of information to shape their efficacy. First is mastery experience, which is the perception that their teaching has been successful through their own experiences. Vicarious experiences are teaching experiences successfully (or unsuccessfully) modeled by someone else. Social persuasion is encouragement and/or specific feedback to teachers about their teaching performance. Last is affective states, where anxiety or excitement, perhaps from receiving results on a recent standardized test, can affect efficacy. Collective efficacy, therefore, is the full faculty believing that they have the power to affect and teach students. Goddard, Hoy, and Hoy (2000) add that teachers analyze the teaching task and assess their collective teaching competence to shape whether they think they will be successful. Collective teacher efficacy leads to teachers more purposefully working to pursue common goals and enhance student learning (Goddard et al., 2000).

### *Challenges to Implementation*

Existing studies explore the various challenges to implementing new initiatives in schools. For example, Roehrig, Kruse, and Kern (2007) use mixed methods to examine 27 high school chemistry teachers as they implemented a new curriculum. They find that teachers' beliefs and preferences for their teaching and the presence or lack of a supportive network within their schools had a strong influence on the implementation. In particular, teachers who primarily used inquiry-based teaching made the transition to the new curriculum more smoothly than teachers who primarily used traditional teaching methods (with instructor-directed lessons that focused on lectures and worksheets), and the most effective support for the new curriculum came from science administrators when they met with teachers to discuss student learning.

Similarly, Briscoe and Peters (1997) qualitatively studied 24 elementary teachers who implemented problem-centered learning in their science classrooms. The primary takeaway from this study is that teachers were best supported in their implementation of the new teaching style by collaboration with their peers. Teachers in the study were able to share useful content and pedagogical information, encourage each other to take risks with the new style, and otherwise support and sustain this change.

Collaboration itself and, particularly, the lesson study approach has been adopted in schools around the world with varying degrees of success. For instance, Bruce, Flynn, and Bennett (2016) look at the efforts of eleven lesson study teams in Canada. The teachers involved taught children aged 4 to 7, and the researchers watched the teacher groups for six months. While there were small student and teacher sample sizes, the research team finds that, because of the lesson study process, teachers' expectations for their children's abilities expanded, and students made significant gains on academic assessments.

Also in Canada, Bruce and Ross (2008) follow 12 teachers as they implemented new teaching strategies and participated in peer coaching. The researchers map teacher change as a function of current instructional practices, student achievement, teacher self-assessment, teacher efficacy, professional development support, peer input, goal setting, and effort expended. While teachers did change their teaching practices to better align with the professional development goals, and teacher efficacy increased, there were also considerable challenges, particularly with peer coaching. Most pairs spanned multiple schools, which made it difficult to meet and maintain productive conversations about teaching. Additionally, the peer coaches were sometimes hesitant to provide substantive feedback to their partner unless that feedback was specifically requested.

Groves, Doig, Vale, and Widjaja (2016) examine the lesson study model in three Australian schools in 2012 to 2014. The collaborative teams were tasked with implementing structured problem-solving lessons, and they found success in deep lesson planning, allowing large numbers of participants to observe their classes, and insight from the "knowledgeable other." However, the teachers had difficulty matching the Japanese problem-solving lesson structure with the prescribed Australian curriculum and had trouble mirroring the Japanese model because of the Australian teaching culture that emphasizes small-group instruction rather than whole-class teaching.

As described by Akiba and Wilkinson (2016), Florida included lesson study in their Race to the Top (RTTT) application, after which the majority of districts mandated lesson study. However, the districts did not first ensure sufficient funding or consider organizational structures or routines in implementation, and only 12 districts (29 percent) requested RTTT funding from the state to support lesson study implementation (primarily for teacher time to participate in lesson study). Because of this, a lack of time and funding was a major challenge for implementation, and schools often used cost-effective substitutes to support lesson study instead of paying teachers more or restructuring to allow for teachers to have more time. Lesson study was shortened and simplified to fit it into the current organizational structures, and it was introduced as one of many professional development activities that occur in schools. All of this increased the quantity of lesson study for RTTT obligations, but it failed to increase the quality, so teachers often did not participate in the research portion of lesson study, examine the curriculum in relation to lessons, or walk through student thought processes while they are learning (Akiba & Wilkinson, 2016).

These examples show how educational practices can be successfully adopted in a new place, but they also show how easy it is to misstep. Imported practices might not be practical for their new environments (Rose, 1991), or implementers might not pay close enough attention to contextualization and ownership of the practices for them to be successful (McDonald, 2012). Akiba and Wilkinson (2016) identify time, a lack of understanding of the research process, and a lack of resources to develop the content and pedagogical content knowledge to maintain lesson study without outside support as major issues when adopting lesson study in the United States. This means that new lesson study models and/or their new environments should be adjusted to account for some or all of these issues prior to implementation so that the practice has a better chance of success.

Alongside adaptations to the collaborative inquiry cycle that happen prior to being introduced to a new environment, it is important to also adapt it to the local environment after implementation, as per my second research question.  McLaughlin (1987) describes implementation “as a process of bargaining or negotiation” (pg. 175), with policies adapting to the local context and the site adapting to the reform, sometimes called mutual adaptation. Adoption of a new practice, like TPEG, should therefore include “additional, individual teacher-directed design, fitting, and adaptation for local circumstances” (Barab & Luehmann, 2003, p. 464) while still being implemented in a way that maintains the integrity of the reform. As an example, Spillane (1999) examined the responses of nine local education agencies (LEAs) to state standards reforms. The LEAs adopted the new standards rather easily but overly adapted the more complex and newer characteristics of the reforms, which led to procedural compliance instead of substantive compliance and change. This meant that the deeper purpose of the reform was neglected, which demonstrates the difference between positive and negative adaptation. Local adaption without losing substance is therefore important to the success and sustainability of new practices.

## Teacher Peer Excellence Groups (TPEG)

The collaborative inquiry model of focus for this study is TPEG. The TPEG model was designed by a group of researchers from multiple American and Chinese universities based on the principles of the Japanese lesson study and Shanghai teaching-study groups (Jensen et al., 2016; Lewis et al., 2006). TPEG was designed to improve on the lesson study model to better suit the American context. The model asks principals or other administrators to implement some of the key characteristics of these lesson study models in their schools. While the administrator leads implementation, makes decisions about the model's evolution, and provides key supports to teachers, TPEG is designed to be led by the teachers so that TPEG teachers feel ownership over the process.

Following the theory of action of collaborative inquiry cycles, the TPEG cycle has four major steps: collaborative lesson planning, observation, lesson feedback and refinement, and storage/sharing. This cycle is completed by teachers in small TPEG teams. In China, teachers are organized into major subject groups that often span two grade levels (Wang, 2013). Similarly, the administrators implementing TPEG often created vertically-aligned teams that focused on either math or reading, which then adapted into subject-grade teams for their daily collaborative practices.

The teams start the TPEG cycle with collaborative lesson planning. Groups of teachers choose the particular concept or lesson to cover and a teacher or two to teach the lesson for others to observe. The teachers pull from resources (including, preferably, a store of lessons that have been used previously) and their own expertise and experiences to plan the lesson collaboratively. This likely includes collective sensemaking, where teachers work together in discussions to interpret and negotiate prescriptive curriculum standards and other policies to include them in their teaching (Coburn, 2001). Wang (2013) notes that teachers should focus on the content of the lesson, for instance, if it aligns with the curriculum standards; how to best teach the content to the students, including discussing what difficulties might arise; and preparing the instructional methods, including any visuals and detailed lists of questions for students.

One or multiple teachers in the TPEG group then models the lesson for the others, meaning s/he teaches it while the others watch, in person or through a recorded video. Importantly, the teachers are observing to evaluate the lesson, not the teacher. The observers can evaluate student reactions, engagement, and understanding better than the teacher who is actively teaching and can therefore provide detailed feedback about the success of the lesson in the feedback session. In China, teachers have official observation notebooks with guides that help document the lesson and enhance their understanding of its effectiveness (Wang, 2013). TPEG groups were given similar guides in their implementation.

The feedback session focuses specifically on the successes and challenges of the lesson and how to best improve it. If every teacher has not yet taught the lesson, the remaining teachers do so with the lesson that has been improved from feedback. This might trigger another round of feedback and refinement of the lesson. After this refinement, teachers are asked to store the improved lesson and accompanying notes in a way that is accessible by other teachers and in future years. For the TPEG pilot, Vanderbilt stored the materials in a central online database that was accessible by other teachers in the pilot.

The TPEG collaboration model was piloted in schools across six districts (18 schools) in Tennessee in the 2013-2014 school year, and nine additional schools were added for the 2014-2015 year. Principals (or another administrator) began implementation at the beginning of the school year then took a trip with the research team a few weeks later to Shanghai, China. The administrator spent a week in Shanghai observing and discussing the local version of the teaching-study groups in a wide variety of schools, after which they were able to return to their own schools and tweak implementation to better reflect the model and their own school culture. Each TPEG school's cycle of lesson planning, observation, and feedback varied from 1 to 6 weeks during the pilot period, and teams submitted documents to the research team after each cycle for analysis.

*Preliminary Evidence on TPEG*

Cravens, Drake, Goldring, and Schuermann (2017) and Cravens and Drake (2017) describe the implementation of TPEG and early results from the pilot using surveys. They find an increase in engagement with deprivatized practice from involvement in TPEG and increased comfort with these deprivatized practices, particularly if the teachers also reported high levels of instructional leadership within their school (principal respected teachers, principal took interest in the professional development of teachers, etc.). Teachers felt that TPEG "provided them with a process for continual professional development, increased the variety of pedagogical strategies they employed, improved student learning, and improved their overall effectiveness" (Cravens et al., 2017, pg. 29). Cravens and colleagues also quantitatively examined other enabling conditions, besides instructional leadership, including teachers' sense of school-wide professional community, teacher trust, and teacher efficacy, and found that teachers who participated in TPEG increasingly engaged in deprivatized practice and collaboration on instruction, even when controlling for those enabling conditions (Cravens et al., 2017). Finally, the researchers found that school context, including student demographics, prior achievement levels, and school size, did not directly predict success with TPEG but that elementary school teachers tended to find TPEG more useful.

These studies provide preliminary quantitative evidence that, given enough support and guidance, teachers can proactively participate in productive learning communities. However, five years after initial TPEG implementation, informal follow-ups by the research team show that there is large variations in how pilot schools engage in these disciplined collaborative inquiry cycles. For the TPEG model to have a more significant impact on instructional improvement, an in-depth exploration is needed to understand how schools address the challenges of time, culture, and resources to implement TPEG.

## Data & Methods

This study seeks to understand the supports and decision-making processes that surround TPEG and teacher collaborative practices. Of the TPEG pilot schools, I did a combination of convenience and purposive sampling to select three schools with relatively close relationships with Vanderbilt that also had various levels of success in maintaining collaborative planning since initial TPEG implementation. At Elwood Elementary School[3], I stratified teachers by grade level, randomly selected two to three teachers to recruit per grade, and then only spoke to teachers who had participated in TPEG. Within Granville Elementary School and Clark Middle School, I spoke to and observed as many administrators and teachers as I could reach in the time I had at each school (two and a half school days at Granville, five school days at Clark). This allowed for stratification by grade, as each team had common plan times during which I interviewed and observed. For most of the interviews, I requested interviews from any teacher I found first during his/her plan time or asked for volunteers to interview after collaborative time ended. There is limited planning time in each school, so I observed as many collaborative sessions as I could and interviewed as many teachers as possible in the time I had available at each school. I spent a shorter period of time at Granville than at Clark and could therefore interview fewer teachers, so I prioritized teachers who had experienced TPEG at Granville over new teachers. At Elwood, four teachers declined to be interviewed, though more did not answer my recruitment email. At Granville, two teachers declined to be interviewed, both because they were too busy, and at Clark, only one teacher declined to be interviewed.

[3] All names of places are pseudonyms.

### *Study Settings*

All three schools are in Tennessee, a midsized state in the southeastern United States. Average test scores of Tennessee students on the National Assessment of Educational Progress are usually average or below average compared to other states, though Tennessee's rankings are improving quickly over time (SCORE, 2019). Tennessee has a variety of state systems and initiatives in education that seek to improve student learning and are relevant to this study, including the Tennessee Educator Acceleration Model (TEAM), the Instructional Partnership Initiative (IPI), and the introduction of new standards. Starting in 2011, TEAM became the primary educator evaluation system in Tennessee, and it uses multiple measures, including observations and student value-added, to evaluate teachers and other educational professionals (TDOE, 2016, 2017). Seventy-six percent of teachers who answered the 2019 state educator survey said that they believe the teacher evaluation process has improved their own practice, and 71 percent said they believe it has improved student learning (*Lessons from Our Educators: Tennessee Educator Survey 2019 Results in Context*, 2019). For IPI, which is a voluntary program that began in 2012, teachers are paired based on their strengths and weaknesses in their teacher evaluation's observation rubric and/or principal recommendation so teachers can observe each other teach, provide feedback, and learn from each other (TDOE, 2017). Finally, in response to the 2015 Every Student Succeeds Act, Tennessee implemented new standards and assessments to be rolled out from 2016 to 2020 (TDOE, 2017). This change required teachers to update their lessons to remain in alignment with the new standards and assessments.

The first Tennessee school, Elwood Elementary School, was particularly successful with sustaining several of its TPEG teams, while others returned to their lesson planning method before the introduction of TPEG (usually individual planning or collaborative planning without reflection and lesson storage) or moved to a new one. Elwood presents an ideal setting for the implementation of TPEG, due to teachers' high levels of collaboration prior to TPEG and the strong support of the principal and academic coach during implementation. However, this school also presents several important challenges to reform sustainability, including the retirement of that principal and coach. The presence of successful TPEG teams, as well as several teams that rejected TPEG after it was initially mandated by administration, means that I will be able to examine both the challenges and enabling conditions of TPEG within a single school. Additionally, some grade-level teams participated voluntarily in a form of collaborative lesson planning, though none maintained the full TPEG process without the administrative mandate.

Elwood is in Cameron County, Tennessee. Cameron County students, as of the 2017-2018 school year, are majority white, and one-third are economically disadvantaged ("State Report Card," n.d.). Elwood was named an award school by the district, prior to interviews, for success with student achievement, closing achievement gaps, and teacher value-added. Table 1 displays additional descriptive information about Elwood and the other two schools in this study. Elwood Elementary was one of the original pilot schools for TPEG. In the 2013-2014 school year, two TPEG teams were started in the school, one for 4th grade teachers, and one for 5th grade teachers. The principal identified these pilot TPEG teams because they were already collaborating frequently. The following school year, the school added a 2nd grade team and a 3rd grade team, and the year after that, they decided to add a kindergarten team and 1st grade team. At first, except during periods of state testing, teams at Elwood met approximately weekly for TPEG, to complete an inquiry cycle lasting two weeks, and administrators and an academic coach attended many of the TPEG meetings. According to the interviews I conducted, all TPEG teams except for the 1st grade team adopted TPEG smoothly, though not all teams still participate or want to participate in collaborative inquiry cycles.

For five years, the official TPEG cycle occurred at different rates each year, sometimes happening only a couple of times total. Part of the decline in TPEG has to do with the retirement of both the principal and an instructional coach who were instrumental in implementing TPEG at Elwood. The new principal at Elwood was supportive of TPEG but was not interested in making it mandatory. Part way through my study, the principal said that her teachers were becoming overwhelmed with an increase in discipline problems in the school and had decided they were no longer interested in participating in either TPEG or my interviews. I have several interviews prior to this point that describe teachers' experiences with TPEG and their daily collaborative practices, as many teams did continue to participate in some form of collaboration. This collaboration included the sharing of lesson planning materials (either during in-person meetings or over the Internet), sometimes a debrief, and the storage of materials. However, the observation portion and the lengthy cycle only continued when teams participated in official TPEG cycles.

Table 1: School Characteristics for Sample Schools

| | **Elwood Elementary School** | **Granville Elementary School** | **Clark Middle School** |
|---|---|---|---|
| **County** | Cameron County | Harrisburg County | Harrisburg County |
| **Urbanicity** | Urban | Rural | Rural |
| **Grades Served** | Kindergarten to 5th grade | Prekindergarten to 4th grade | 5th to 8th grade |
| **Approximate # Teachers** | 50 teachers | 20 teachers | 35 teachers |
| **Began TPEG** | 2013-2014 | 2014-2015 | 2014-2015 |
| **Mandatory Collaboration** | No | Yes | Yes |
| **Approximate # Students** | 1000 students | 400 students | 800 students |
| **Approximate Student Demographics** | 85 percent white,<br>10 percent economically disadvantaged | 90 percent white,<br>30 percent economically disadvantaged | 90 percent white,<br>20 percent economically disadvantaged |

The second school, Granville Elementary School, had less success with sustaining the original TPEG model, but more success sustaining their evolved collaborative practices. Granville Elementary significantly contributes to this study because it had required collaboration, instead of voluntary, as at Elwood Elementary. Granville is an interesting case because it regularly uses IPI, the peer observation initiative. IPI has overlap with TPEG, so many Granville teachers had a difficult time differentiating TPEG from IPI before I reminded them of the full scope of TPEG. Granville continuously participated in IPI since its introduction, and its focus was on teachers observing each other and providing feedback. While these observations were not connected to specific lessons, as they were with TPEG, it is informative to examine why this observation-focused model has been fully sustained at Granville for several years, whereas Granville teachers ceased observations connected with their collaborative lesson planning, as was introduced with TPEG.

Granville Elementary started TPEG in 2014-2015 with two teams that were vertically aligned, one teacher from each grade comprising a TPEG team that was kindergarten to 2nd grade and another team with teachers from the 3rd through 5th grades. Teams did TPEG cycles approximately once per semester for about two years. Neither the teachers nor the administration were able to identify in interviews why official TPEG cycles were dropped. They were able to speak at length, though, about their current collaborative practices. The introduction of collaborative planning at Granville was not met with much resistance by the teachers. The teachers spent this time planning lessons or sharing materials, grading together, debriefing on challenging lessons, and storing materials either on paper or digitally. Granville Elementary was in Harrisburg County, where the students, as of the 2017-2018 school year, were majority white, one-fifth were economically disadvantaged, and had very low populations of

English language learners and students with disabilities. The economically disadvantaged students at this school did not test well, but their growth scores were high ("State Report Card," n.d.).

The third Tennessee school, Clark Middle School, was also in Harrisburg County and was particularly successful with collaborative planning schoolwide. The school frequently hosted educators from across the state to observe and ask questions about their collaborative practices. Teachers at Clark ceased incorporating every part of the TPEG process in their daily practice, but every major-subject teacher collaboratively planned every lesson. Like Granville, Clark had mandatory collaboration. The teachers at Clark had a significant amount of time to collaborate and stuck particularly closely to the original TPEG theory of action for their daily practice, leaving out only the observation portion. While the original TPEG teams were formed on a voluntary basis, the introduction of mandatory collaboration was met with some initial resistance, which then went away as teachers either left the school or, more often, saw how impactful the collaboration was to their teaching practice and time management. As of my interviews and observations, teachers at Clark maintained collaborative inquiry cycles that were the most closely aligned with TPEG and were the most uniformly enthusiastic about their collaborative practices, compared to teachers at the other two schools.

Clark Middle School started TPEG in the second year of implementation, spearheaded by an assistant principal. Two subject-specific TPEG teams (one math and one ELA) were formed the first year. In the spring of the second year of TPEG, the principal and an academic coach attended a TPEG meeting and got the idea to expand TPEG so it affected more students at once. They decided to restructure the master schedule so that each major subject-grade team had 45 minutes of mandatory collaborative planning time every day and an additional 45 minutes of flexible plan time per day. The science and social studies teachers tended to teach the same lessons for both A and B days (Clark has a block schedule), so they had an hour and a half of mandatory collaborative planning time every other day. These subject-grade collaborative teams planned every lesson together in person. The specials teachers, like art and STEM, did not formally have collaborative planning time but had common plan times with other specials teachers and were informally encouraged to collaborate. After this new schedule was implemented, the teachers no longer participated in formal TPEG cycles, and their new collaborative practices did not include observations.

### *Data Collection & Analysis*

Data collection took place from the spring of 2018 to the fall of 2019. I first collected phone interviews from seven teachers at Elwood Elementary, as shown in Table 2, before they opted out of continuing participation in the study. The principal explained that this was due to discipline problems that prevented TPEG from continuing and made teachers feel they were too busy to be interviewed. While I opted for in-person interviews at Granville Elementary and Clark Middle, I believe the phone interviews at Elwood gave me extra insight into the relationship between years of experience and views on collaboration that I would not have gotten with in-person interviews due to my young appearance. Participants at all three schools were aware that I was a graduate student, so those on the phone might have guessed I was young as well. I was able to complete data collection in-person at Granville and Clark. At Granville, I conducted 14 interviews and five formal observations. At Clark, I conducted 27 interviews and 10 formal observations. Interviews were transcribed by me or the service Matchless Transcriptions.

Table 2: Statistics on the Qualitative Data Collected

| | **Elwood Elementary School** | **Granville Elementary School** | **Clark Middle School** |
|---|---|---|---|
| **# Administrators Interviewed** | 0 | 2 (100% of total) | 3 (100% of total) |
| **# Major Subject Teachers Interviewed** | 7 (~15%) | 10 (~50%) | 23 (~75%) |
| **# Special Area Teachers Interviewed** | 0 | 2 (~50%) | 1 (~10%) |
| **Median Experience in the Education Profession of Participants** | 10 years | 15 years | 14 years |
| **Average Interview Length** | 40 minutes | 13 minutes | 12 minutes |
| **# of Plan Time Observations** | 0 | 5 | 10 |

Interviews from the three schools lasted from 4 to 55 minutes, depending on the amount of time the respondent had available, and the observations were approximately 30 minutes each. I was initially concerned that I would not have enough time to build trust with respondents to get honest answers in the shorter interviews, which were done exclusively at Clark Middle School, but almost all of the teachers seemed very comfortable speaking with me and responded in similar ways as those with whom I spoke for a longer period of time. This (non)reaction could be because the teachers at Clark saw me around the school for multiple days in the halls and observing their collaborative sessions, and they were used to having visitors observing and asking questions about collaborative planning at least a few times per year. Additionally, their journey with collaborative and deprivatized lesson planning likely made many of them feel confident speaking publicly about their practices.

In the interviews, I asked about issues that are difficult or impossible to directly observe, such as trust and participants' views on collaboration, and those that require retrospective accounts, such as the implementation and adaptation of TPEG at their school. The interview protocol (see Appendix A) was designed to ask about views on TPEG, specifically, and collaboration and professional development more broadly. First, I asked about teacher background and teaching style to get demographic information and some basic information about TPEG participation. I used that information to determine which sections to cover, using a skip pattern. The skip pattern allowed me to ask questions that best fit the participant, depending on if they had used TPEG or collaborative inquiry cycles for several years, if they still participated in TPEG/collaboration at the time of the interview, if their team used to but had ceased to participate in TPEG/collaboration, or if the participant had never participated in TPEG/collaboration.

I had different interview guides for teachers and administrators/coaches. In the teacher interviews, the sections that referenced collaborative inquiry, particularly for participants that had been participating in it for several years, focused on how teachers were trained to participate, the structure of the cycles and its evolution over time, their and their colleagues' reactions to it, enabling conditions, and how they judged whether a lesson was successful. For instance, when I moved from a discussion of the early days of collaborative inquiry to the current day, I said, "Let's fast-forward to today and talk about what, if anything, has changed with TPEG/collaborative planning since your earlier interactions with it. What does a TPEG/collaborative planning cycle look like now? Has the overarching structure changed much over time?" I then asked participants to describe a cycle, including who attended meetings, how the team chose lessons on which to collaborate, and when teams met with each other. To inquire about enabling conditions, I first asked an open question, "What was it that made TPEG/collaborative planning work or not work well for you and your team back then?" then followed up with specific examples from the literature review and other participants to inquire about their relative importance. If I had remaining time in the interview, I also asked broadly about collaboration and professional learning. In particular, I wanted to learn the participants' ideal collaborative environment and professional learning opportunity.

I had a separate interview guide for the administrators and instructional coaches. While the structure of the interview was the same, including the skip pattern, I asked these participants more detailed information about implementing TPEG in their school, key stakeholders, and decision making on how to adapt TPEG over time. I asked specific questions about each school's resources and decisions about how to schedule collaboration and who should and should not participate. Each school had one or two administrators/coaches who were deeply involved in the adoption and adaptation of TPEG, including traveling to Shanghai, changing the master schedule, and supporting the collaborative inquiry. The assistant principal at Clark Middle seemed to think of her interview as a way to help others learn of her school's tremendous success with collaboration. The principal at Granville Elementary was attempting to practice what she preached; she said that she wanted to be interviewed because she encouraged her teachers to make their practice public, so she should make her practice as an administrator public as well.

While I began and ended with the same questions for every participant, I used the interview protocols as a loose guide for the conversation. In particular, I altered the order of questions frequently to help respondents flow from one topic to the next naturally. I wrapped up each interview with an open question about if there was anything else I should know about the participant, their school, or collaborative inquiry, and many participants took this time to speak warmly about their colleagues, administrators, and/or students. As I went through the interview process, I updated the interview guide as needed, primarily to get more specific information on current non-TPEG collaborative practices.

For the observations, I went to collaborative team planning sessions. I did not use a formal observation protocol, but I focused my notes on time use and the type and quality of interactions between the teachers on each collaborative team and any visitors from the administration. The observations were approximately 30 minutes each. Most collaborative teams used their collaborative time as usual, though a few stopped their work to talk with me about this study. The latter teams sometimes joked about specific moments that I took notes. When asked, I explained that I was there to observe the structure of their time and interactions, rather than judge their performance or any lack of focus, which seemed to put the teams at ease and allow them to commence work without letting my presence distract or alter their collaborative time.

Analysis was conducted using the software Dedoose using line-by-line coding and hierarchical coding for inductive discovery of details that are salient to the participants and deductive coding based on specific aspects of the TPEG conceptual framework. In particular, I first identified enabling conditions that are named in the literature review, how to promote buy-in, other initiatives, and support for collaborative inquiry that allowed it be continually successful for some teachers and not others. Next, I examined each step of the TPEG cycle and how it adapted over time to see which aspects of the TPEG cycles seemed to best suit and support teachers in these contexts. Third, I looked at teacher instructional practices and instructional improvement, as expressed by the teachers and administrators, to determine whether the existing adaptations to the TPEG model are still contributing to fulfilling a professional knowledge base, as per Hiebert et al. (2002).

## Results

*RQ1: Why have some schools or some teams of teachers continued to use collaborative inquiry cycles while others have stopped?*

"We had some genuine dislike [of mandatory collaboration] that was just small here and there…So we really tried to look at that and pitch a positive note to it…just mandatorily require it, and then once they got together, they saw, hey, this is not so bad." - Administrator PC

Sustainability is an important concept in education, as many initiatives and practices are introduced and discarded on a yearly basis in schools. The above administrator explains that her school grew a successful and sustainable collaborative environment by marketing collaboration to teachers as a positive change, making it mandatory, and supporting teachers while they made sense of the change and found ways to make it work. The actual process was, of course, more nuanced, but there are lessons from these schools, particularly in regards to instructional leadership, about introducing and sustaining collaborative practices that can be described and analyzed in context to help other schools considering instating collaborative inquiry cycles.

*Forming Collaborative Teams.* In the three schools, the majority of the effort put forth by administration took place when TPEG was introduced and when collaborative practices were scaled up to the entire school. The first decision that needs to be made is how to group teachers into collaborative teams. Collaborative teams in Shanghai often share a major subject but span multiple grades (Wang, 2013), so many of the TPEG teams were formed similarly, as can be seen in Table 3. At Granville Elementary, teachers formed vertical TPEG teams that were not subject-specific, even though a few grade levels were departmentalized. Teachers at Clark Middle formed vertically-aligned TPEG teams with other teachers in their same subject. At Elwood Elementary, TPEG teams were more flexible, so teachers usually collaborated with their grade level but would occasionally collaborate with the grade above or below them to improve the school's vertical alignment. Although many teachers in this study acknowledged that vertical alignment could be useful or interesting, those benefits quickly got overshadowed by time constraints, and many teachers found it frustrating to spend so much time helping other teachers plan a lesson that they would never teach themselves. TPEG teams tended to be large – about six people at Clark, up to eight at Elwood, and two or three at Clark – but the sustained collaborative teams were smaller with two to four people in each team. The actual size of the collaborative teams was not particularly salient to participants, so larger teams might be appropriate in schools with more teachers per grade or subject-grade.

Table 3: Collaborative Team Groupings

| | **Elwood Elementary School** | **Granville Elementary School** | **Clark Middle School** |
|---|---|---|---|
| **Initial TPEG Teams** | Whole grade with some vertical | Vertical (K-2, 3-4) | Vertical by subject |
| **Sustainable Collaborative Teams** | Whole grade with some grade-subject and some same-subject vertical | Whole grade with some grade-subject | Grade-subject with some same-grade, cross-subject as needed |
| **Plan Time** | 45 minutes per day | 50 minutes per day | 90 minutes per day |
| **Mandatory Collaborative Time for Lesson Planning** | N/A | 25 minutes 4x per week | 45 minutes per day |

The shared grade or subject-grade teaching responsibilities among collaborative team members were salient and important to the teachers. One Clark Middle School teacher gave an example of why vertical teams did not work well:

> For example, the first [TPEG] one we did was a 5th grade English lesson. It didn't seem all that applicable to many of us, because they were focusing so much on fluency and basic comprehension and parts of their standards that we don't even have those sorts of standards in [grades] 6, 7, 8. So I remember that that was– You really felt like you were helping one other person's lesson or that 5th grade group's lesson, but you didn't feel like it necessarily applied to you. – Teacher LC

This is also some evidence that collaboration worked better with teams that were not as familiar with the curriculum standards. At Granville Elementary, teachers and the principal reported that some of the best collaborative teams formed when teachers in the team were all new to the subject-grade. This seemed to be because the teachers all had to create completely new lessons, so they did not harbor as many protective feelings over their existing materials. They felt more willing to split up the work and share their newly-created materials with each other than if they had a store of materials already individually prepared.

*Scheduling Collaborative Time.* After deciding on how to form collaborative teams, administration needs to set up plan time so that teachers on collaborative teams have opportunities to plan together. For TPEG, this time was usually after school, a practice that was not popular and did not last for long. The sustained collaboration in these three schools occurred in the day during embedded plan periods. Elwood Elementary School participated in TPEG for the longest period of time, and their TPEG meetings occurred several times throughout the semester during professional learning community (PLC) meeting times. This PLC time was embedded in teachers' days and specifically designed for collaborative activities, which is why it was relatively easy to maintain TPEG practices.

TPEG lasted a shorter period of time at Granville Elementary and Clark Middle. At Granville, TPEG teams met after school, and the sustained collaborative teams met two to five times per week during their daily 50-minute plan periods. At Clark, the administration gave teachers incentives to stay after school to participate in TPEG. After about a year and a half of this, the administration decided to implement daily mandatory collaborative lesson planning that has been sustainable at the school. They rearranged the schedule so that every major-subject teacher got 90 minutes of plan time every school day, the first half of which must be dedicated to collaboration.

All three schools have this grade-level plan time to support grade or grade-subject collaborative teams. However, there are not always at least two teachers teaching the same grade-subject, so sometimes administration had to get creative with the collaborative teams, as was the case at Clark Middle. Because the common plan time was grade-level specific, if there was only one teacher who taught a major subject in a specific grade, s/he would be placed with teachers who taught as similar a subject as possible. In one grade, the single science and single social studies teachers formed a collaborative group. In another, the single science teacher joined with two math teachers. In these collaborative groups, the discussions were necessarily less specific. Teachers were able to engage in sensemaking and logistical conversations and would exchange ideas, rather than debrief on or plan specific lessons. While each teacher who joined these cross-subject collaborative teams expressed their great value, many also shared that they would have preferred to have been matched with someone in the same subject.

Time is always a concern in schools, but the administrators at Clark Middle and Granville Elementary were able to increase the overall plan time to accommodate collaboration and individual planning needs without increasing costs. Both looked carefully at their schedules to see where they could be made more efficient and cut minutes from parts of the school day that were less necessary (like long hallway transitions). While Clark is not a Title I school, both of its feeder elementary schools are, and the school does not bring in much money, according to the assistant principal. She discussed the tradeoffs inherent to ensuring that teachers have enough collaborative and plan time, namely that administrators were able to give their teachers 90 minutes of daily plan time by increasing class sizes to an average of 35 students per class and decreasing the amount of plan time special area teachers have.

The administration was able to convince special area teachers to accept this arrangement because of the close relationship administration has with all of the teachers in the school and by emphasizing that special area teachers' evaluations are based on school test scores, and major-subject teachers' collaboration would increase school test scores.

Sustainable collaborative teams at these three schools had grade or grade-subject groupings with common plan times embedded in the day for the team. Collaborative groups having a common plan time works well for lesson planning, feedback and lesson refinement, and storing lessons, but it does not allow for teachers on a collaborative team to easily observe each other teach as a data-gathering mission. I go into more depth about observations in a later section, but a major problem with TPEG observations was that teachers did not want to leave their own classrooms to observe another teacher, even if they had appropriate coverage for the class. All three schools stopped using observations to get feedback on their lessons, so I cannot comment on existing solutions to this issue. A possible way to retain the observations that could be explored further is to have differently scheduled plan times during the week, perhaps three days of collaborative time with the grade and two days vertically with the subject, so that teachers can collaborate during common plan times and observe each other during uncommon plan times. This varied scheduling could also help ensure that teachers get some individual planning time, which teachers expressed they needed. As one teacher said, “None of us necessarily would want to get to the point where we're using our plan time every day to meet with each other because we want to have time in our classrooms to get other things done. Even just to put copies away or look through kids' papers” (Teacher DE).

Given limitations on time, it is important to carefully choose how frequently teachers should collaborate. Granville Elementary, Clark Middle, and some teams at Elwood Elementary participate in daily collaborative planning, but for TPEG, teachers were only asked to focus on one lesson approximately every two weeks. An administrator from Clark explained why her school decided to shorten the inquiry cycles:

> The two week [TPEG] cycle that was vertically aligned, and they work just fine, but they weren't an everyday fix for an American school…it took so long, we weren't getting enough bang for our buck, I guess is how you could put it…it was one lesson for two weeks for one grade level, one subject area to be impacted…90 kids maybe, that are going to be impacted by it. So it wasn't the maximum impact we could have. Because of the size of our building and the number of teachers we have and the way we were able to arrange our schedule, it worked out where we could impact every single kid every day. So the payoff to that was much better than doing a two-week cycle. – Administrator PC

This daily cycle has clear benefits, including that it affects every student every day, but it also means that teachers are not easily able to incorporate parts of the research process into their cycles. The collaborative lesson planning becomes part of the daily teaching practice, rather than a distinct activity that uses research techniques to deliberately focus on creating and testing particularly high-quality lessons. Some teachers suggested that the “formal” TPEG cycle that focuses on one lesson over the course of several weeks could be incorporated into their school again but that there should be at least a couple of weeks in between each cycle. Over the course of a few years at Elwood Elementary and Granville Elementary, however, those short breaks between cycles devolved into several months and then eventually the extinction of TPEG in both schools.

*Learning to Collaborate.* Once collaborative teams are formed and given time to collaborate and perhaps time to observe each other, teachers need to learn how to collaborate productively, a skill that does not come naturally to many people. Administrators that led TPEG attended information sessions that taught them best practices in collaboration, and one administrator from each pilot school went to Shanghai to observe lesson-study in practice in a variety of Chinese schools. These administrators were then tasked with explaining and modeling proper collaborative techniques at home and ensuring that their teachers had or were building a professional community and trust amongst themselves and were gaining ownership over the TPEG process. At Clark Middle School, the administration did this by showing videos from the trip to Shanghai, modeling good collaborative practices in front of and with collaborative teams, sharing research on collaboration, and providing a simple checklist of how to perform productive collaboration. As one administrator lists, “The very first thing is to review the lesson they’d just taught. What was good, what was bad, what needs to be changed, and then where do we need to go from here, and that’s when today’s lesson [planning] begins.” After a while those steps became internalized to the point where most teachers at Clark forgot the checklist existed.

This internalization is important because one of the most salient (and negative) aspects of TPEG for teachers was all the paperwork that was involved in the pilot. Vanderbilt requested that TPEG teams upload detailed lesson plans and reflection guides that felt burdensome and time-consuming to most teachers. Veteran teachers at Elwood Elementary, in particular, felt that the requested lesson plans were much more detailed than they needed to successfully teach a lesson, so creating the detailed lesson plan took out more time from their days than was necessary. Teachers at Elwood Elementary and Clark Middle both had central systems where they have to upload their daily teaching materials, but they are not more detailed than what the teacher him/herself used. At Granville Elementary School, teachers did not have a central system to store materials, so they locally saved only necessary materials on hard drives or in filing cabinets to access the following year.

*Expectations for Collaboration.* Alongside explaining and modeling good collaborative practices, the administration cultivates a culture of high expectations for collaboration to maintain fidelity and improve rewards from the collaboration, both of which increase teacher buy-in and ownership over the process. Clark Middle School is the perfect example of this because the principal maintained extremely high expectations for fidelity to his school-based teacher collaboration routine when collaboration was first scaled up to the entire school, and that eventually resulted in a very strong culture of productive collaboration. The principal explained,

> They had an hour and a half planning period. The first 45 minutes had to be co-planning, and there were no exceptions to that. Not going to the copy machine. Not having IEP meetings. Not going to get a snack out of the [vending machine] thing. This is co-planning time…If your expectation is that they will be doing this for 45 minutes, and if they're in the hallway, that it is addressed very quickly and that there's no doubt about what they're supposed to be doing for that 45 minutes. And then if you do that a couple of times, everybody has [it]. – Administrator OC

During data collection at Clark, many teams started their collaborative planning quickly in the plan period, followed the steps the administration outlined (what went well, what went poorly, what needed to improve), and stayed longer than the 45 minutes to plan together. The mandatory collaborative time seemed to be a way to protect teachers' time so that they are forced to do something that they know is good for them and their teaching (collaborative planning) but is easy to leave behind in the shuffle of all of the various tasks teachers are expected to do. While many teachers and the administration said the 45 minutes of collaborative time was "sacred," I found it to be *mostly* sacred with occasional interruptions for children to visit the classroom and the principal to make team announcements. This was apparently good enough, as observed collaborative teams tended to use all of their time to reflect and create high-quality lessons. A Clark administrator recounted a story about an observer from another school:

> Someone came and observed us one time at one of the schools that came in, and I never will forget this. She made a comment to [a teacher], "Wow, it's like you're planning to be observed every single day." And [the teacher] was like, "Uh huh." Like, that was a foreign statement to her. She didn't really understand – I mean, she understood why she was saying it, but like anybody in our building would go, "yeah." (laughs) Like, of course we

> are. Shouldn’t you be? Like, you’re teaching every day. It should be your top-quality lesson, you know? So the [visiting] teacher was just flabbergasted. – Administrator PC

At Granville Elementary, the expectations were not quite so strict, and the resulting collaborative practices were less cohesive. However, the administration at Granville was committed to strengthening expectations around collaboration every year, which will presumably lead to even more closely aligned planning and teaching. One administrator explained, “Through the years we've tried to give our teachers that expectation that we're all on the same page. You can have a variety in how you personally teach, but as far as the curriculum and the expectations of everybody learning things at the same pace, just every year we've tried to make that stronger” (Administrator GG).

Elwood Elementary School is a contrasting case in that, at the time of data collection, some teams did collaborate while others did not. There was no longer an expectation that teams collaborate, which allowed me to examine other enabling conditions for their relationship with collaborative practices. According to teachers who had not been on the team when collaboration was mandatory, the 1st grade team did particularly poorly with TPEG and did not sustain most collaborative practices after it was no longer mandatory. Teachers reported that the 1st grade team was less cohesive than the teachers in other grades, and one teacher who was formerly on the 1st grade team described her team as “segregated.” The evidence suggests that the 1st grade team lacked the professional community and trust that enable successful collaborative practices.

Mandatory TPEG, or other collaboration, forces teachers to take the “extra” step of collaborating, which is something many would like to do. While some teams clearly found intrinsic reasons to sustain collaboration, there are so many ways that collaboration can be derailed or found to be too much work, that the mandatory nature, along with attention to buy-in and the usefulness of the practice, seems key to the sustainability of the practice. This might seem obvious, that making something mandatory makes it more likely to be sustained, but I think it is an important point to make. It would not be surprising if the administration at Clark Middle and Granville Elementary eventually felt that their school had a strong collaborative environment and no longer needed such firm expectations around their collaborative practices. Giving teachers more flexibility might allow them to better shape their collaborative practices around their own needs. However, if this were to happen, I would expect at least some teams at both schools to scale back their collaboration and return to individual planning, as happened to teams at Elwood Elementary.

Teachers at all three schools related how they were initially unhappy with the mandatory collaboration. A teacher and instructional coach at Clark explained how it was received by teachers:

> I think one thing [that made collaboration work] is that our principal made it non-negotiable. And like at first, he'd be in the hallway, and if we were at the copy machine or going to the restroom, he would yell at us and tell us to get back in our room, that this was collaborative planning. So even for people who were real reticent to do it and who were like, "Well, I'll just go sit in someone's room for 45 minutes, but I'm not teaching what they're going to teach," the fact that they had to stay in a room and pretend they were collaborating, eventually they did collaborate. And no one was like thinking, "I don't like this person" or "I don't want to spend the time." Where they were thinking was, "My students aren't like her students. My classes aren't the same. I don't teach like her." We had one group of two teachers who were both very strong but one is very tech heavy, and tech savvy, and the other one was almost afraid of technology and never used it, and both of them came to me separately and said, "No. This is not going to work. I'm just going to pretend. I like this person, but I'm not going to teach like this person taught." Within a year they were absolutely pacing together, teaching the same lessons. I think that they both learned what the other person can bring to the table, so that was a big piece is kind of making it non-negotiable, and it's just part of our culture [now]. – Teacher/Coach LC

Over time, these teachers and many others were able to see that the benefits of collaboration outweighed the problems. These teachers found that there was more flexibility in the practices than they originally thought and grew to love the collaboration. It was during this time that administrators also transferred collaboration from an administration-led initiative to a teacher-led initiative. An administrator at Granville Elementary explains that she and her vice principal "used to go in every week and just sit through planning times, but the more times we were in there, the quieter they got, and it got to be a session that was driven by us, not by them. So we don't go in every single week" (Administrator GG). The same happened at Clark Middle School. The school leadership had to gauge how to maintain high expectations for collaboration and offer support while remaining mostly separate from the practice of collaboration.

*Buy-In.* In addition to instructional leadership, trust, efficacy, and professional community were important in sustaining collaborative inquiry cycles, primarily through the importance of establishing buy-in. Participant comments show that it takes time to cultivate buy-in for the collaboration, and both administrators and teachers needed to be convinced that it would work. An administrator at Clark Middle School said that her trip to Shanghai was "vital" to prove to her that it would work, and it took about a year for her teachers to see the fruits of their labor (high-quality lessons stored in a central location) and be convinced that the collaboration was a good idea. At Granville Elementary School, one administrator explained how she needed to spark the interest of just a handful of teachers to cultivate enough buy-in for the collaborative teams to use best practices in collaboration. The administration at all three schools tried to spark interest, relate stories to their teachers, and roll out collaboration slowly to promote buy-in, but their eventual success came from teachers seeing improvements in their own practice and workload over time after administration made TPEG or the evolved collaborative practices mandatory.

Buy-in also developed through the efforts and patience of the teachers themselves. The teachers at Clark Middle and Granville Elementary seem to have strong shared norms about student learning being at the center of their practice and about how they have a collective responsibility for all of the children in their grade. Teachers at Elwood Elementary who did not like collaboration spoke frequently about their preferred teaching style, rather than about practices that would most help their students learn. One veteran Elwood teacher who no longer collaboratively plans talked about her grade-level team:

> We…have a great difference in teaching styles. They, in general, [do] more of a PowerPoint lesson and some other things that go with that, and that's not a natural way of teaching for me. I would rather just talk or do group activities or do some inquiry or do some lecture or do different things. And my way of teaching is much different than the ones on my grade level at lots of times. I like the way I do it, and they enjoy the way they do it. And so I don't really care to spend a lot of hours trying to prepare something that involves, you know, a certain way of presenting the material…There is a big difference in the way that [my teammates] teach in lots of ways. And I'm not saying that's bad. I'm saying that it's wonderful and accomplishes the purpose. But it's a kid of a different method that I don't care for sometimes…When I've taught this for so many years, it's not that I don't want different ways or new ways. I just like to take the best of every aspect that I can find and then make it what I want it to be. – Teacher BE

A veteran teacher at Clark was similarly resistant to the introduction of mandatory collaboration, but he was required to continue using it and eventually grew to appreciate it. He explained:

> I'm an old guy. I don't like change. (laughs) That's the reality of it. I've always just been resistant to change. I get into what is comfortable for me and then something comes along and wants to realign the parameters of it, and usually I'm not happy with it. And my experience typically is even though I make the change begrudgingly, it winds up being a positive thing, and that's been the situation [with collaboration]…How we accomplish that [collaborative lesson] in the classroom, it's not identical. And I don't think you can have an expectation that it would be, and I may be out of line with that, but you have two different personalities involved. And for me, the critical element in any classroom is the teacher and how my partner teacher creates the atmosphere. Where she communicates what she wants to communicate and how I do that, we can do it two separate ways and

> still be effective. If we share what our content ambition is and then we arrive at that product in a different way, I don't have a problem with that. – Teacher YC

This mindset came more often from teachers who had been in the classroom for a long time, and several administrators and younger teachers expressed that they thought the more senior teachers had the hardest time adjusting to collaboration. Teachers at all three schools, though, expressed how strongly they all contributed to their professional communities. One teacher said,

> When I'm here at this school, it's different because we're working as a team looking at all of our data collectively. So we want everybody to grow. It's not, "I'm collecting my own data from my own 1st grade classroom or 3rd grade classroom and saying this is how much I've grown." Yes, we pull individual scores, but we also look at how did our 5th grade department. And that little shift in looking at ourselves as a whole and making sure, okay, we got this, we got this…I think setting first the parameter of the kids are our main focus. We want them to grow, giving each other praise on things, not taking individual praise for something that occurs. So if we try an idea, we all try it. This isn't my idea. We're doing this together. – Teacher JC

Her principal expanded on the idea:

> Here in this building, we have a lot of very strong professionals who do what it takes to get the job done. Do what it takes to be the best. You don't have that everywhere. So it could be more difficult…We have to try to weed them out. Some people are happy just not working hard every day because it's a lot of hard work, you know, to plan and prep high-quality lessons every day. It's not easy. – Administrator OC

Individual efficacy was not very salient with the teachers and administrators in this study, even when they recalled their teaching practices prior to collaboration. One way that individual efficacy was mentioned was with one teacher who wanted a grade-subject collaborative partner (instead of a same-grade, different-subject partner) to check his work. He says, "My thing is to keep myself honest, you know, like in that if I'm doing something that – to truly know how effective it is, I need someone else really to do the same thing, you know?" (Teacher CC).

The teachers on collaborative teams at all three schools had high degrees of collective efficacy. An administrator at Granville Elementary describes her school:

> This [school] is where I feel good, because we have excellence here, but you also have a heart too. So, you know, like, our kids that you've seen that are struggling, you know, we don't just stick them in a room somewhere. You know, we've got people, and you saw

> folks who are – my pre-K, that was a teaching assistant who came to help with a little girl in first grade because she had time. The teacher she works with had her kids in specials, so she had a little time and she could have just kind of hidden out in her room and said I've got things to do, but she came on board and tried to help out. So everybody – my RTI teacher, she had a few minutes, so she was down – and she was saying, where can I help, how can I help? And that's the culture of our school. So I think that's what keeps us strong. – Administrator GG

What was most prevalent, though, were discussions about trust in teammates. This manifested both in trusting that teammates were going to produce high-quality work and that teachers felt they could be vulnerable in front of their teammates to make their teaching public. One Elwood Elementary teacher explained, “There’s a good level of trust insofar as level of excellence. You know, there’s no one that stands out doing an absolute wonderful job while there are some others that are just along for the ride, that doesn’t happen” (Teacher BE). An administrator described the fears of some of her teachers:

> It’s scary…if I open my planning time to you, and you’re the other teacher coming in, and we’re going to plan together, what if your ideas aren’t as good as mine? Like it’s that fear. Or what if you haven’t taught at all and I’ve taught for 25 years, you know? It’s that fear. So it’s overcoming that and really shifting the mindset from type A personality, I have total control, because a lot of teachers have that personality (laughs) to, okay, let’s put our minds together because we’re both adults, and we’re both professionals, and we both have degrees, both very well educated. You have strengths, I have strengths. Let’s combine those, and let’s work on each other’s weaknesses, and I think that we are at a point for that now. It takes the time to get there, though, for sure. – Administrator PC

This administrator detailed a mindset shift that happens over time. Teachers talked about it similarly, as a shift that came either after a working together successfully for a while or a conscious decision to trust each other with their shared work. In these three schools, though, teachers on collaborative teams were only able to give feedback on lesson planning and anything they were able to glean from their debrief. They could not comment on specific teaching because they did not observe each other teach. The dynamics of trust between team members would likely be altered if that component were left into the inquiry cycles. As it was, though, teachers could filter or color their reports of how lessons went when debriefing with peers.

Because of how important trust was to respondents, I was surprised that teachers did not seem phased by turnover on their teams, particularly at Clark Middle School, where teachers shifted frequently between grades, subjects, and even schools. While teachers admitted that there was an adjustment period with a new partner, they seemed to take a “mind over matter” approach to quickly adapt to new teammates. One other factor that might have been contributing to this ease was that teachers at Clark were frequently observed by teachers and administrators from other schools (about once every two months) who wanted to learn about their collaboration, so they were used to their practice being public to a wide range of observers.

Important to collaboration is that teachers feel comfortable disagreeing with each other in order to create high-quality work. In collaborative sessions, teachers provided positive feedback to each other frequently, but they also discussed and negotiated points with each other, particularly in one team that considered themselves to be "highly productive." A teacher from that team said,

> People's feelings get hurt sometimes because you kind of go in with a whole lot of ideas…Tomorrow we're going to have to decide [what will be in the lesson], and we can only probably do a third of the things that we talked about doing, and some of us will win, and some of us won't, and that'll be okay. We're pretty used to it, but sometimes people's feelings do get hurt because you really have strong opinions about how it should be. Sometimes the solution to that is let's do something different in this section…You know, maybe we're doing a little action research right in the middle of the lesson. And sometimes the answer is a debate. You know, and then we decide which one we're going to do. – Teacher LC

This teacher was implying that they can disagree with each other productively because they had practice in resolving disagreements. Another teacher from that team referenced their mutual respect and communication skills for resolving disagreements.

Sometimes disagreements are not built on this shared commitment to collaboratively improve teaching. Many teachers in the study identified interpersonal skills as the most important factor in creating productive collaborative partnerships. The other side of that is that they also identified personality clashes as the easiest way to disrupt collaborative relationships. In particular, personality clashes tended to happen when teachers on a collaborative team had very different teaching styles or when one teacher had a much bigger personality than the other and attempted to control the decision making in a way that was unwelcome.[4] Sometimes, the mandatory collaboration forced teachers to learn how to work together productively, usually by realizing how they could combine their strengths, but sometimes, an administrator or other neutral party with some authority, like an instructional coach, mediated the relationship. They would help members of the collaborative team align their goals and

[4] Some collaborative partnerships were like mentorships with veteran teachers doing the majority of the decision-making when their partner was new to teaching and/or the grade-subject. These relationships were usually welcomed by everyone on the collaborative team and were not seen as disruptive.

priorities and made sure that everyone was following best practices with collaboration. One teacher described this process:

> My second year…was a bit difficult of a time. It was just kind of a mess, and we had to have some instructional coach come and help…Things weren't working for [my partner] with admin, and so it just kind of got onto me. So it was just back-and-forth checking how much I was doing but then checking to see if she was doing similar stuff…[Having the coach step in] was very helpful. She helped guide us on what needed to be done and, you know, she said you've got to do this, so it helped kind of cut everything, like, you got to do this, you got to do this, so that was very helpful. – Teacher ZC

Occasionally, if this mediation did not work, administration would coach a resistant teacher out of the school. This apparently happened rarely, and I did not get specific information about it, but it does raise questions about whether the seemingly necessary high expectations for collaboration would break down in schools that could not afford to lose a teacher because of their lack of interest in collaborating.

After a while, all of this effort that was put into establishing a structure and expectations and supporting the collaborative teams culminated into a strong collaborative environment. At Clark Middle School, the assistant principal said that after the first year, her teachers were able to pull from their bank of lessons that were collaboratively planned and stored in a central system, meaning they only had to slightly update the lessons before teaching them, which convinced them of the importance of collaboration. This appeared to be the key, at least for teachers at Clark, to get complete buy-in for this type of collaborative practice. Teachers said that getting used to intense collaborative relationships could be uncomfortable, particularly if teachers observed each other and/or gave feedback on lessons, but it was often worth the trouble. At Clark, teachers had been participating in intense collaboration for about four years when I visited the school, and almost every single teacher passionately expressed how valuable collaboration is. They reported that it improved their workload and their teaching, as evidenced by improvements in their teaching evaluations and their students' standardized test scores. For schools that also do observations, like Granville Elementary (though their observations are unrelated to lesson planning), it can take a long time to get used being observed and giving constructive feedback, and Granville's principal reported that it took three to four years for the teachers to start looking forward to being observed and critiqued by their peers.

*RQ2: Compared to its original theory of action, how has TPEG evolved to the local context?*

The TPEG process has four major steps: lesson planning, observation, reflection/refinement, and storing/sharing. While none of the three schools in this study specifically continued to use TPEG in its original form, teachers at each described their prior interactions with the TPEG model and how they still participated in most of the steps of TPEG for their daily lesson planning, just less formally or perhaps not every time they planned a lesson.

*Lesson Planning.* The lesson planning process for TPEG varied between schools. At Granville Elementary, TPEG teachers would observe and reflect on an individually-planned lesson, which is not true to the TPEG model. At Elwood Elementary and Clark Middle, two to three people on larger TPEG teams would collaboratively plan a lesson then share it with the rest of their teammates. Early on at Elwood, TPEG teachers would collaboratively plan whole-group (teacher-directed) instruction and focus on a particular indicator of their new TEAM teacher evaluation rubric. Several teachers found this to be helpful to get to know the new rubric, though after a while they appreciated having more flexibility in what to focus on when collaborating. For instance, a few months before her interview, one Elwood teacher had used TPEG collaborative time to teach about and model for her peers an intensive reading program that she had learned about in a professional development course. Some teachers reported feeling like they had to make perfect lessons for TPEG and therefore spent an exorbitant amount of time on lesson planning when it was their turn to create the lesson. Teachers also described how frustratingly detailed their TPEG lesson plans were required to be to submit them to Vanderbilt.

Once TPEG faded in the schools and smaller collaborative groups that met daily or almost daily emerged, collaborative lesson planning became much more focused on the daily needs of lesson planning. Some teams, like those teaching in Granville's lower grades, have very strict county standards and even activities, so much of their collaborative time involved using sensemaking to understand and organize the materials from their county. Almost every teacher reported starting the lesson planning process by examining their materials from previous years. They then usually made minor updates, based on how the lesson went last year, which was either written down or remembered. Many teachers updated the lessons based on their individual group of children in the current year, but others would only update the lesson if it went poorly the previous year, not if it went well.

Teachers described how they sometimes had to make more substantial updates to the stored lessons or remake them completely. This usually happened because of a curriculum standards change or because the teachers wanted to theme the lesson for a holiday. The teachers at Clark Middle felt a high expectation for excellence and completely remade their lessons more frequently than teachers at the other two schools. One Clark teacher said her team creates entirely new lessons about once every two weeks to make sure they are pushing themselves to teach at a high standard. Another teacher explained her team's willingness to create new lessons and bring in new materials because of their own love of learning. Clark teachers had an advantage when updating lessons based on a standards change because they had access to the materials from all teachers in the school. If a standard moved from one grade to another, which happened frequently, then Clark teachers could easily access the materials of the teachers who taught that standard before.

Many Clark Middle teachers stayed close to the original lesson study model by using their lesson planning time to anticipate student questions and difficulties and how they might overcome them as instructors. During one planning session, a veteran teacher at Clark prepared a lesson with her partner, who was a novice teacher. After making a PowerPoint together, the veteran teacher explained exactly where students might get confused. The veteran teacher helped the novice teacher clarify her own understanding of the content and then provided examples of exact wording and hand signals that the veteran teacher thought students would find helpful.

That example is one of many that illustrates how teachers thought of collaboration and were best able to use it to improve their instruction. Teachers described building off each other's experiences, pushing each other to try new techniques, encouraging each other to see problems from new perspectives, and generally having multiple minds working together to improve their teaching. While these were most relevant for teachers who taught the same content, it also applied to cross-subject or cross-grade collaborations, where teachers could bounce ideas off each other and push each other to think and teach in new ways.

The collaborative teams had preferred collaborative styles, each with different strengths and weaknesses. I identified three major styles. The first way of collaborating is what I call *planning together*, where the teachers meet to plan lessons at the same time. Some teachers might already have ideas of what to put in the lesson, but often those ideas are fielded as possibilities that the collaborative team then narrows down to put into identical or nearly-identical lessons. This is the traditional lesson study method of planning, and many collaborative teams at Clark Middle, Granville Elementary, and Elwood Elementary use this style of collaboration. It has the distinct advantage that each lesson is created by the group of teachers who can combine their knowledge and experiences to make an excellent lesson.

Another collaborative style is *sharing lessons*, where teachers split the work into distinct units that are individually planned and prepared to share with peers, who often do not change anything about that lesson before teaching it. One teacher described the process for her team:

> We actually set aside each day. Like I might have Tuesday/Thursday lessons, another teacher will have Monday/Wednesday lessons, and one other teacher will have the Friday lesson or a test that she creates. So, every week, I know I've got two lessons that I need to make, and they're going to be awesome…If it's an activity that requires worksheets or any kind of supplies, I make sure all of my colleagues have those things. So all they have to do is show up and teach it. – Teacher TC

Each lesson is prepared individually, so the lessons themselves do not benefit from collaborative thought, unless the team decides to change something when the teacher is sharing his/her portion of the lesson with the other teachers. However, it allows each individual teacher to have more time to devote to his/her portion of the lesson and other responsibilities, including those outside of work. The teacher quoted above felt that her lessons were improving over time, largely because each teacher on her team made positive updates to the lessons stored from prior years. In this way, the improvements to the lessons are what one would get if working alone while adequately storing and referencing materials year to year except that teachers involved in this method of collaboration have more time to spend on the lesson plans because they only have to work on a portion of the total lessons.

The third style of collaborative lesson planning is what I call *sharing materials*. With this method of collaboration, teachers have a connection (likely digital, though it could be in person) where they share ideas and techniques with each other that they are not *expected* to use. One Elwood Elementary teacher attributed her preference for this collaborative style due to her many years of experience teaching and her comfort with her own teaching style. She described this collaboration as a way to get new ideas instead of being boxed into another teacher's style of teaching:

> I love to get copies of [my teammates'] notes. I love to just share what I'm doing with them, and if they don't want to do it that way, that's just fine with me…It's not that I don't want different ways or new ways. I just like to take the best of every aspect that I can find and then make it what I want it to be. Rather than everybody agree to say this and this and this and use this worksheet and do these notes on an active board. Some of that I love. But I don't care for being, kind of, molded into this exact way of doing it…I probably sound like I just want to go off on my own with no collaboration and no teamwork, but that's not the case at all. I do love to share and love to gain different ideas from other people, but I want to pick which ones I want to use and which ones I don't. – Teacher BE

Some people might not consider this collaboration, and it certainly is not the type of collaboration that is envisioned to connect with lesson study. However, some teachers, particularly those who do not like the idea of teaching the exact same lesson as his/her peers, prefer this method. Other teachers can *only* use this method due to staffing or structural constraints. At Clark Middle School, teachers who do not have a grade-subject collaborative partner often have to use this method of lesson planning. They might work together to solve problems, but much of their actual lesson planning is individual with some degree of sharing techniques or resources that can span subjects with each other. Vertical collaborative teams would have to collaborate similarly, sharing techniques and resources that can span grades.

Every collaborative team that was observed or interviewed used one of these three styles of collaboration or a combination. For instance, sometimes teachers would collaboratively plan with the first two styles. In that case, the substance of the lesson would be planned together and then the work of actually preparing the materials would be split up (one person updates the PowerPoint from last year, another prepares the materials for an activity, etc.). That way, the lessons still have an overarching collaborative vision, though specific details are decided on an individual basis. Or, they might collaborate on one portion of the lesson and individually plan another, like if a team collaboratively prepared the part of the lesson that was preplanned by the county, so they could participate in sensemaking activities together, but individually planned the rest of the lesson.

Which collaborative method is best? Lessons that are planned together likely become the highest quality lessons, but it can be difficult for teams to develop and maintain the advanced communication skills that are required to allow for this style. Sharing lessons requires trust in teammates' abilities, which also takes time and effort to develop, but it seems to result in higher quality lessons over time and allows for teachers to spend less time on lesson planning overall. Sharing materials might also allow for higher quality lessons over time, but it is very teacher-dependent, as it allows teachers to maintain much of their privacy and does not have added pressure for teachers to plan great lessons because they will all be teaching it. However, sharing materials could be a good transitional collaborative style for those who are resistant to collaboration.

*Observation.* If the lesson materials are planned to be the same across classes, then delivery of the lessons is logically going to be similar as well. In all three schools, teachers who planned for the lessons to be the same described how similar the delivery actually was. Several Clark Middle teachers said that 85 to 90 percent of their lessons would be the same across rooms, though some teachers would switch the order of activities, put personal touches on the lesson, like a favorite PowerPoint background, and adjust the difficulty of lessons based on the level of his/her students. At Granville Elementary School, many teachers asserted that their whole class instruction would be nearly identical across classrooms, particularly in grades that had strict county materials to follow, but that small group instruction and perhaps warmup materials would vary from room to room. Interestingly, all the teachers at these two schools would describe the similarities between their classrooms and the differences – due to differences in teaching preferences and students' abilities/interests – in much the same way, though some described these differences as massive, while others described them as important to their individual identity as teachers but small.

Observations of the lessons is the part of collaboration that was most salient, discussed, and fretted about by the teachers in this study. At Granville, some teachers were perfectly fine being observed and would not change his/her practice due to an observation, particularly an informal observation by a peer teacher. Most teachers, however, became nervous when "tall people" came into their rooms. Many of these teachers reflected that it is easy to try to put on a show or get fancier in their teaching when someone is observing. One teacher explained how she always had good experiences with observations and getting feedback but that it was still a nerve-wracking process. She said,

> I would much rather be in the company of seven- and eight-year-olds than grown-ups because I feel like they're more forgiving. If I make a big goof with an adult you think, "Oh gosh! I know they probably caught that." Where with children, if you make a mistake, you're like, "Oops, nope, that's not what I meant," and they're like "It's okay, it's good." You know that they're being truthful because that's just how they think. – Teacher JG

Quite a few participants had similar stories. Teachers realized that it was silly that they were nervous having adult observers but were anyway.

Teachers also struggled with exactly how they were supposed to conduct the observations. With lesson study and TPEG, teachers are supposed to pay attention to the lesson and how it translated into practice, rather than paying attention specifically to the teacher. Said another way, they are supposed to evaluate the lesson that was collaboratively planned, not evaluate the teacher him/herself. Some (but not all) teachers who participated in TPEG seemed to understand that they were supposed to make that distinction. It helped teachers who were nervous about observations begin to rationalize their way into accepting the observations. However, even those who understood the distinction had a difficult time figuring out how to make it work in practice. For the most part, teachers described observations as a time to watch teaching practices. They tended to not pay special attention to student reactions or student work. Only one teacher explicitly mentioned how in-person observations could be helpful to observe student engagement. Another teacher explained how observations were only necessary to get to know her teammates better:

> I just don't think [observations] needed to continue after a couple of cycles just because once you see someone teach and you know that person, then you're familiar, you have the background information, you don't really need to see it all the time because you know that teacher is still going to be a strong teacher and how they perform…At the beginning I felt like [observing each other] was a big part to it, and we benefited from the strengths of other teachers because it was an eye-opener to see how other teachers taught the same lesson that we had. – Teacher GE

She went on to describe how the observations were also great for what teachers could see and gain beyond their shared lesson plan, like ideas to improve classroom management.

In fact, these peripheral things, beyond the lesson, seemed to be what the teachers in the study preferred to take from observations. I identified two main reasons that teachers at Granville Elementary School liked doing these observations. The first was to find tips and tricks that they could bring back to their own classroom. Often these were related to classroom management and motivating students or activities that seemed engaging. The second reason was so that teachers could better understand their vertical alignment in the school. For this, teachers tended to prefer observing classes that were one grade ahead of his/her own class. That way, s/he could see what his/her students would need in the future and what expectations their new teachers would have. Teachers felt they could orient themselves by watching how other teachers interacted with students in class.

A few years after TPEG was introduced at Granville Elementary, the administration brought in a practice to improve vertical alignment called Feedback-by-Teacher (FBT), which is an extension of the state Instructional Partnership Initiative (IPI). With FBT at Granville, teachers were paired to observe each other teach for about 20 minutes and then give feedback once each before they moved to another teacher. They did this process approximately twice per semester. The principal at Granville introduced this practice slowly, first just having teachers observe each other and learn anything they felt they could. After a while, she introduced a guide that corresponds to the teacher evaluation's observation rubric for teachers to follow in observations. The principal said it took about three years before teachers became excited about participating in FBT. Almost every teacher at Granville that participated in TPEG got TPEG and FBT confused with each other because they were introduced around the same time, and the observation part of TPEG seemed to be the most different from teachers' daily collaborative practices.

At Elwood Elementary and Clark Middle, observations were dropped completely when TPEG evolved into other collaborative practices. Most of the Elwood teachers interviewed vividly remembered teaching and observing TPEG lessons. One problem they encountered was that they did not like when another teacher who was covering their class was in charge of teaching the TPEG lesson. To solve this problem, the TPEG teams worked with the principal and an academic coach to change the schedule so each teacher could teach her own TPEG lesson, which helped to alleviate the problem.

Another major problem for observations, for both TPEG and FBT, was finding the time to do them. At each school, teachers of the same grade level shared a plan time, so they could only observe teachers at a different grade level if they were going to observe during that time. This not a large problem for teachers at Clark, who had 45 minutes of individual plan time every day. At Granville, however, plan time is very limited, so frequent observations significantly detracted from teachers' tolerance of the practice. Administration provided substitutes for TPEG so teachers could observe each other when they otherwise would be teaching a class, though most teachers did not like leaving their students.

Each school discussed or experimented with using technology to ease the burden of observations. Granville teachers talked about videoing themselves teaching TPEG lessons, but they did not feel they had the equipment or expertise to do that well, so they only did in-person observations. For a short time, Clark Middle School teachers watched observation videos as a group immediately before debriefing on the lessons, but they preferred doing the in-person observations. At Elwood Elementary, teachers found it difficult to get virtual observations right. Even though they were able to figure out the technology, they found it to be a "big load," particularly with finding the time to watch the videos. There were also some teachers who felt very uncomfortable being videoed. For all of these various reasons, digital observations were not sustained for long at the three schools.

Peer observations of any kind did not last in any school as part of the lesson planning process. Teachers routinely talked about observations as something "extra," which decreased their motivation to participate. One teacher said, "I work really hard during the day and try to get everything done. I don't always necessarily have time for the *extra* of it" (Teacher AE, emphasis mine). Another teacher in the same school elaborated, "You can't fully understand the amount that is on your plate, so then just to add even an extra little meeting or having to make coverage plans that often and that kind of thing. It can just become a lot" (Teacher EE). Interestingly, teachers at all three schools expressed their vast appreciation of observations in their interviews. Almost no one wanted to get or give feedback based on observations, but most enjoyed observing one another teach. It seems like mandatory TPEG, or FBT, forces teachers to take that “extra” but beneficial step of observing each other.

The problem with collaborative teams abandoning the observations is that they are vitally important to the collaborative inquiry process. They are the way for teachers to evaluate the lesson by capturing whether students are engaged and actually learning the material. Because of this, each collaborative team needed to find new ways of collecting information on the lessons.

At Clark Middle School, teachers evaluated lessons by paying attention to their own impressions of the lesson, including overheard student comments, and almost obsessively examining student assessment data. To afford so much plan time, classes average 35 students, which is about five students more than before the current collaborative practices started. Despite this, teachers believed they were able to adequately pay attention to student engagement, perhaps because students at Clark seemed to feel comfortable speaking up frequently and honestly in class and were friendly with their teachers outside of class. Teachers reported that when they debriefed on lessons, one person’s comments often remind the others of what went well and poorly in their own lesson. This method of remembering and recounting is practical, though it allows for a degree of privacy in that teachers’ recollections are filtered through their own memory and desire to share.

What Clark teachers paid the most attention to, though, was student assessment data. One part of their teaching that Clark teachers were required to standardize across classrooms was their student assessments. Teachers frequently made students complete exit tickets and, at least once a week, a short quiz. With assessments that are standardized across classes, teachers are less able to maintain their privacy and could evaluate the strength of their lessons based on how well the students are able to demonstrate their knowledge. This, of course, rests on the presumption that exit tickets and quizzes are an appropriate way to capture student learning and, as an extension, how well the teacher taught. If one is confident in this, then this method gives teachers a way to evaluate a small number of lessons at a time, which teachers can then use to update those lessons when they are taught in the future.

As the younger elementary school students are less able to express themselves and take frequent assessments, Granville Elementary teachers had to find different ways of evaluating their lessons. For individual lessons, teachers reported watching students to see if they were “glazed out.” They would also pay attention to whether students would correctly use new information, for instance, a new vocabulary word, at a later time outside the context of assignments or assessments. In the longer term, teachers tracked goals for their classroom and individual students, paid attention to their teacher evaluations and their students’ standardized test score growth, and talked to teachers in the grade above to see if their students were adequately prepared. While these might help a teacher evaluate if s/he was a good teacher, most of these methods are not useful for evaluating individual lessons. The only method that Granville teachers could routinely rely on to provide information about each lesson was the teachers' memory of how the lesson went, which, again, is filtered first by the teachers themselves.

*Reflection & Lesson Refinement.* At Granville Elementary, teachers used their recollections to debrief casually on many, but not all, of the lessons. Teachers caught up in the hallways between classes or at lunch to ask about how lessons were going before meeting for their formal collaborative time. Based on observations, their critiques of the lessons were often based on whether the teacher liked the lesson and its delivery, rather than framing discussions around how much they thought the students learned from it. If the lesson seemed to go well, they might not give it more than a cursory discussion. Granville teachers reported that sometimes they would write notes to store with the lessons based on their debrief to better guide them the following year in lesson planning, particularly if a lesson went poorly, but usually their discussions were only for immediate use.

In one collaborative session, the assistant principal came in to update teachers on school activities. I took observation notes on the collaboration when he left:

> *One teacher comments about needing to prep a reading activity, and another takes her to the correct book and talks about how she is going to incorporate it into her own lesson. One teacher is updating an activity from last year and wants to go print the materials. She offers to print it for the others then leaves to do that. Two teachers stay in the classroom to talk about how they aren't sure why kids missed a certain question on the assignment and their scaffolding strategies. They aren't talking about the kids' thought processes, though, why they picked that answer. They talk about how there are some things they tell their kids that they just have to memorize without knowing why they work because they're tricky topics…Their support for each other seems to be mostly about sensemaking and getting advice on new ideas.* – Granville Elementary Observation Notes, Day 2

What struck me about these interactions was how fragmented they were, with pairs of teachers breaking off to discuss or share, and how the interactions were often focused on assisting each other, rather than on building something together. The two teachers who discussed the assignment seemed to use their conversation to better support their students' future learning, but it is unclear whether the other two teachers who left the room got those benefits as well. The teachers in the discussion also did not seek out additional resources when they were not quickly able to understand their students' reasoning.

Many of the teams at Granville seemed to rely on each other for emotional support. In one collaborative session, teachers graded assignments with each other and would occasionally lift up a paper to show others. While grading, they would share stories, usually funny or frustrating ones, from their classes. Some of these stories were to prompt a discussion about classroom management or a teaching technique, but many of them seemed to largely be about gaining emotional support. Another team used the start of their collaborative time to talk about an intruder drill that was scheduled to happen soon. The teachers spent several minutes expressing their fears about the drill and a possible intruder. They also practiced some sensemaking, figuring out the logistics of the drill and how it would affect their week before sharing funny stories from their classes and giving each other advice on how to handle certain children or lessons that were coming up. Little (1990) calls this storytelling and scanning for ideas, and she regards it primarily as a method for teachers to reveal their knowledge, intentions, and values to his/her peers and shape or reinforce their shared professional community. As in the example above, it might also shift into what Little (1990) calls aid and assistance, where colleagues assist in the practice of teaching only when asked and avoid giving unwarranted advice on the stories. This appeared to be the boundary between emotional support and professional support in the observation at Clark Middle School that I describe below; teachers in this Granville team often sympathized about the plight of their peers in the stories but only gave advice in moments where the teller was clearly seeking additional support.

The collaborative teams at Clark Middle School, despite having more overall time dedicated to collaborative planning, stayed more focused during lesson debriefs on whether and how their students learned and gave each other more direct professional support. In one collaborative session, the collaborative team was focused on the results of a quiz they had given the prior week. Each teacher had the quiz results up on personal computers and a room projector, though it was clear they both had examined their own class's data before the meeting. They compared how quickly their students completed the quiz and went over almost every question together. If there were discrepancies between the classes, the teachers would compare exactly what they taught and how they taught it, bringing up particular comments or discussion questions that the collaborative group had not discussed before teaching the lesson. Other debriefing sessions were similar, with teachers comparing student assessment or assignment data, discussing particular questions that several students had gotten wrong, why they likely got them wrong, and how the teachers could adjust their next lesson to clarify any misconceptions.

Like observations, teachers reported that feedback sessions can be nerve-wracking for some. With TPEG, only one teacher was usually observed teaching the lesson before the lesson refinement, so the debrief could feel like an attack on him/her personally. Even without observations, teachers might feel the need to keep information private or talk themselves up to colleagues to maintain their image as a competent teacher. At Clark Middle, this did not seem to be an issue. Teachers sounded frank when describing how the lessons went and sharing how their students did on assessments. Likely, over time, these teammates grew to trust each other with an honest performance review. This is possibly because the teachers being able to present their own evidence and self-reflection to the group about whether their lesson was successful (rather than teachers gathering evidence on each other's lessons) allowed for them to feel open about discussing their weaknesses and analyzing the differences between classes. There was even some friendly competition between teachers, which one described as her way to keep pushing herself to improve her teaching. Almost every team gave off a strong sense of collective responsibility for all of the students in their grade, whether or not the students were in their own class.

Many of the Clark teachers interviewed said that they always immediately update lessons that need it, and sometimes teachers would also include a note about their changes. These changes were immediately helpful for the handful of teachers who taught the same lessons to different students from one day to the next, but many of the teachers did this simply for their own benefit in the following year. Despite this level of detail of discussion and updating lessons, Clark collaborative teams rarely spent more than the first 10 minutes of their planning period debriefing on and refining a lesson. Over half of their collaborative time was dedicated to planning upcoming lessons.

Elwood Elementary School teachers were able to share specific information about TPEG reflections, and each teacher seemed to have a different experience with the feedback. One teacher exclaimed that she was able to use reflections to brag on her peers about what went well in the "model" TPEG lessons and what everyone could learn from the observation. Another felt that TPEG encouraged everyone to self-reflect in a way that she found helpful. Some teachers felt that TPEG reflections were a good way to get "safe, comfortable" feedback in a way that was less formal than the observations that accompanied their teacher evaluations. However, others felt that they had to give surface-level or even biased feedback to avoid hurting people's feelings. Some Granville Elementary teachers felt this as well in their FBT observation feedback and expressed that it was not helpful to spend time on the FBT process if they could not give or receive substantive feedback. Based on Granville and Clark Middle in particular, giving constructive feedback seems to be a skill that can be learned, though, so perhaps more time and training would alleviate some of these issues. Also, these particular concerns are almost exclusively connected to giving feedback after a peer observation, rather than debriefing after teaching a lesson that peers did not observe.

*Storing/Sharing.* With TPEG, lesson storage and sharing is supposed happen frequently throughout the process. Teachers pull lessons from storage when they are lesson planning, store lessons before teaching them, and store the updated lesson after the debrief and refinement. However, storage and sharing were not salient to many of the teachers in this study, and many only commented on it when prompted or when they shared about the challenges of the TPEG pilot. Quite a few teachers identified filling out detailed lesson plans and storing materials online for Vanderbilt as the single most challenging part of TPEG. As the pilot progressed, teachers were given more leeway in how they approached TPEG, including with paperwork, which Elwood Elementary teachers in particular welcomed with relief.

At all three schools, it was clear early on that the TPEG pilot storage system was not going to be sustainable, so they switched to their own systems. Teachers at Elwood had storage online that allowed them to share materials with each other and the principal for comments and access state curriculum standards. At Clark Middle School, the county set up a central lesson repository for storing and sharing materials across the district. Clark teachers were required to store all materials for every lesson in this easy-to-use online system. The Clark administration did not require teachers to submit daily lesson plans, though, because they found it was not useful for most Clark teachers. Administrators occasionally gave feedback on stored lessons, and some teachers reported accessing other grades' materials to stay informed about vertical alignment and to pull materials when standards changed grades.

Teachers at Granville Elementary School had many possible ways to store lesson materials, so the particular method ended up being specific to each team and largely localized. Lessons were often kept on someone's hard drive and/or in a filing cabinet. Some collaborative teams stored materials online in Dropbox or Google Drive. Like at Clark, Granville had access to an online system hosted by the county with materials from teachers in other schools, but teachers were not required to use it. Several teachers mentioned occasionally downloading materials from this county system, but no one described uploading their own materials. The Granville storage methods were therefore often used only as convenient "storage units," rather than as extended spaces for collaboration or the sharing of materials, even among team members.

*Non-Negotiables.* Describing how TPEG evolved begs the question of whether these new collaborative practices fulfill the non-negotiables for communities of practice that should translate a practitioner knowledge base to a professional knowledge base. The three non-negotiables are that the practice of teaching is public; there is a mechanism to validate and improve the lessons; and the improvements to teaching from the collaboration are cumulative, accessible, and shareable (Hiebert et al., 2002).

Starting with public, Hiebert, Gallimore, and Stigler (2002) first state that knowledge "must be represented in such a way that it can be communicated among colleagues" (pg. 7). They then used examples of real people to say that,

> The insights [the teacher] acquired…will not contribute to the profession's knowledge until they are made public and examined by others. In a sense, what Grace learned was public because she shared it with Stephanie; they could describe and understand what they were learning. But professional knowledge must be created with the *intent* of public examination, with the goal of making it shareable among teachers, open for discussion, verification, and refutation or modification (pg. 7).

Did the teachers in this study reach an appropriate level of publicness to contribute to the professional knowledge base? Those who collaborated certainly had intent to share their lesson materials with each other, though those who collaborated via sharing materials did not open their creations up to be discussed, verified, or refuted. Those who collaborated via planning together and sharing lessons, however, did. Even if they rarely refuted lessons that were shared with them from colleagues, they were usually given the opportunity to do so and could discuss the lesson in more depth in their reflection after teaching the lesson. Using this definition, the teachers who planned together and shared lessons do seem to be adequately making their teaching public.

Second is whether there is a way to validate and improve the lessons. At first glance, there is. Teachers at Clark Middle, in particular, spent time dissecting their own impressions and student assessment data to reflect on their teaching and improve lessons, particularly if the lessons were perceived to go poorly. The question becomes more complicated when considering whether the teachers are able to adequately reflect on their teaching, given that they did not observe each other teach the lessons. Lewis, Perry, and Murata (2006) consider live observations as critical to lesson study as part of the research process. Perhaps American teachers are more able to accurately judge student reactions than Japanese or Chinese teachers, as American class sizes are smaller on average, by up to twenty students per class (OECD, 2018; USDOE, n.d.). Are teacher impressions of his/her own class enough to validate and improve lessons? What about student assessment data?

Additionally, Hiebert, Gallimore, and Stigler (2002) differentiate between local knowledge generated by the teachers themselves, which might not always be accurate, and expert knowledge or repeated evaluation in different contexts. Expert knowledge comes from instructional experts such as coaches, some administrators, and researchers. At Clark Middle School, the instructional coach spent half the day teaching, and her collaborative partners expressed their appreciation of having her on their team to share expertise. All three schools had instructional coaches, but their roles were usually to provide assistance based on requests from the teachers. Sometimes coaches and administrators would sit in on collaborative sessions to share expertise, but, particularly at Granville Elementary, they only did so rarely to ensure that teachers maintained ownership over their collaborative practices. Additionally, no interview participant described researchers who were instructional experts being involved in the lesson planning process. At Clark Middle and Granville Elementary, teachers had access to materials from other teachers in the county (repeated evaluation), but they mostly relied on their own materials and only referenced county materials or those from other grades when confronted with a new curriculum standard.

What does this mean? Given the daily collaborative inquiry cycles that most teachers in this study used, there was not time or interest in daily observations to collect data for validation and improvement. Experts could be involved in these daily cycles, likely with some increased expense to the school or district, but almost none of the collaborative teams in this study used experts. Granville Elementary and Clark Middle teachers could, but did not, learn from repeated evaluation of their lessons throughout the district, though it would require increased coordination with other teachers in the county. Because of these three points, the teachers in these schools primarily relied on local knowledge, which means that they were not guaranteed to be appropriately validating and improving their lessons.

Finally, the product of this validation and improvement must be storable and shareable. Hiebert, Gallimore, and Stigler (2002) assert that it is not enough to share locally with a few colleagues. Professional knowledge must reach beyond the time and place they were created. Elwood Elementary has only a school-wide system, but Clark Middle and Granville Elementary have online county systems. Again, Granville teachers do not share their materials on the county system, but they could, perhaps if they were required to do so. These online repositories allow the lesson materials to reach more teachers than they otherwise would have. Is it enough that the lessons are only shared within the county, though? Harrisburg County, where both Granville and Clark are located, had distinct curriculum standards that were updated every few years. The teachers at these schools closely examined curriculum standards when planning new lessons, which means the lessons were closely aligned with county expectations. Perhaps this means that most lessons should only reach so far as the county.

Places that defer instead to other district, state, or even federal standards might want to extend their lesson storage system to those levels instead.

*Improvements to Practice.* The theory of action behind lesson study is that teachers collaborate by examining and improving lessons together to make themselves higher quality teachers so their students get better instruction. While there is evidence that lessons and instruction are improving, which improves student outcomes, it is unclear whether teachers themselves are improving. The simple fact that they are collaborating and sharing high-quality lessons means they will be improving some. However, imagine a situation in which two teachers are collaborating, one novice and one veteran. The veteran teacher brings his/her many years of experience in the profession to their collaboration, and the novice teacher brings his/her knowledge of new techniques and technologies. If neither teacher goes into depth about *why* their contributions are important or *how* to best incorporate them into future lessons, then the two teachers might eventually separate and use their new knowledge with lesser effect than when the two collaborated. Have these teachers become stronger because of their collaboration or were they only stronger in the context of their relationship or the exact lessons that they planned together?

This is not a hypothetical problem. Morris & Hiebert (2011) discuss the Japanese lesson study and emphasize that stored lesson plans include rationale for teaching decisions and changes, so that other teachers can later use the lessons in new contexts. One administrator in this study shared that she thought about this potential problem with her newer teachers, who had only ever known collaborative planning. She worried that they might leave her school and be unable to plan new high-quality lessons on their own. Perhaps if teachers in every school collaborated, this would not be a problem. That is not the case in most places, however, so teachers who transfer schools would likely need to know how to plan independently.

Despite the intermediary, the ultimate goal for lesson study models is to improve instruction. Many teachers and administrators stated that collaboration made them grow as teachers by allowing them to learn from their peers, forcing them to detail their thought processes when planning lessons, holding them accountable to each other to do high-quality work, and often decreasing total work load. The administration at Clark Middle School, though, expended effort to track concrete changes:

> One, we were a rewards school this year. Our overall student [testing] data has gone up…We have also seen [improvements] in teacher overall observation scores on the TEAM [evaluation] rubric…[The principal] gets evaluated every year just on his ability as an administrator, and he has seen a rise in his scores in this. We have seen a rise in

> happiness ratings [from approval surveys] from our teachers. And one thing that is always on the evaluation is, "Please don't stop our collaborative plan." – Administrator PC

This administrator, and teachers at both Clark Middle School and Granville Elementary School, attributed the increases in student test scores, teacher evaluations, the principal evaluation, and teacher approval ratings directly to their collaborative practices. Collaboration also allowed for teachers to participate in sensemaking and emotional support activities in what is usually an isolating career path.

## Conclusion

"The American teaching culture, I think, is very different [from that in Shanghai] in the fact that teachers are not natural sharers. We have always really taught with a closed-door mindset" (Administrator PC). Teachers in three Tennessee schools have been able to overcome this mindset to become collaborative partners to improve their teaching. I came away with five major lessons, mainly about instructional leadership, about how to enable collaborative inquiry cycles in these schools:

1) Form subject-grade collaborative teams,
2) Create time for collaborative teams to meet that is embedded within the school day,
3) Get buy-in through proof that the collaborative inquiry cycles will be worth the commitment, from observing other schools with successful collaborative practices and/or seeing changes in teachers' own practices,
4) Instruct and/or model how to productively collaborate, with particular emphasis on how to give constructive feedback, and
5) Maintain high expectations to participate in collaborative planning (perhaps by making it mandatory) and to create high-quality lessons, among both administration and the teachers themselves.

I also identified five ways that teachers and administrators in these schools decided to adapt or implement the inquiry cycles into forms that are more sustainable and/or successful:

1) Teachers tend to collaboratively plan lessons multiple times per week, instead of spending multiple weeks examining one lesson,
2) Planning lessons together, instead of other collaboration styles, appears to create the greatest benefit for both teacher and student learning,
3) Peer observations should be conducted in-person and scheduled so that teachers do not have to leave their students,

4) Thorough, though not necessarily lengthy, reflections will ideally be conducted on every lesson, and

5) Online spaces that are shared among the collaborative team, the administration, and preferably, teachers outside the collaborative team are better to store lesson plans and reflection notes than are more localized storage units.

There are a few pieces collaborative teams in these three schools did less successfully that future research should address. First, many of these teachers struggled to appropriately gather data to reflect on the success of their lessons. When they observed each other, they did not fully understand what to observe, and the model teacher often felt uncomfortable in the process. When they did not observe each other, they used a variety of techniques to gather information about their lessons, but it is unclear how objective and meaningful those sources were. Second, it is still unclear how to move from local knowledge from the teachers themselves to incorporating expert knowledge to better validate and improve lessons. Principals or other administrators could participate in collaborative meetings as instructional experts, but it is often important for them to maintain their distance to ensure that teachers feel ownership over the collaborative inquiry cycles. One possible solution would be to embed instructional coaches, perhaps one who also teaches the same content, so they are seen as insiders, into collaborative teams. Other questions that this research opens up are about how to help schools transition from noncollaborative to highly collaborative environments. For instance, can collaborative teams use sharing materials and/or sharing lessons to ease the transition from individual planning into the more intensive method of planning together?

This study is distinct from others already in the literature because of its focus on decision-making and tradeoffs when administrators and teachers implement and adapt collaborative inquiry cycles in American schools. It also helps fill some of the "critical research needs" that Lewis, Perry, and Murata (2006) identify. In particular, it describes a lesson study practice called TPEG and how it was supported and evolved over time in three schools, each with distinct characteristics and ways of practicing lessons study. It also helps explain the mechanism by which lesson study can improve instruction by improving lesson plans, which is why teachers in my sample decided to collaborate on every lesson instead of focusing on one lesson for several weeks at a time. I also provide evidence that it improves teachers' commitment and community, though there is less evidence that the collaboration in this study improves teachers' knowledge or learning resources.

This research has several limitations. One is the interruption of data collection. At Elwood Elementary, I wanted to complete many more interviews, including ones with the former principal and instructional coach who were instrumental in introducing and support TPEG at Elwood, and do observations. However, I was asked to cease data collection at Elwood after having completed only seven phone interviews. Some of the plan sessions I observed at Granville Elementary and Clark Middle were abnormal, as the Granville assistant principal needed to make announcements to each team on one day that I visited the school, and it was the end of a grading period at Clark, which meant that the plan sessions were more casual than usual. Additionally, some of the teachers had difficulty remembering the early days of TPEG or differentiating it from the FBT observation initiative. I was able to find at least a few of people at each school who remembered TPEG vividly, though, so I felt confident in reporting their descriptions of and reflections on that time.

Another major limitation is external validity. I can identify and analyze patterns and processes in Elwood Elementary School, Granville Elementary School, and Clark Middle School, but the findings will not directly translate to other schools. In particular, these were all middle- or high-achieving schools prior to the introduction of TPEG, and I was asked not to follow up with teachers at Elwood when their collaboration was reduced due to an increase in discipline problems. However, I expect many of the decisions and thought processes around balancing resources and desiring improved instruction will be consistent across many schools.

The teaching culture in the United States is very private and has a variety of other challenges that oppose the use of collaborative inquiry cycles as a method to improve student learning and teacher working conditions. While this study does not prove that TPEG or lesson study conclusively benefits education, it does describe how three Tennessee schools navigated the process of making teaching more public and of higher quality. Along with other studies, this research can help inform how to best implement and sustain teacher collaborative practices in schools.

# REFERENCES

Achinstein, B. (2002). Conflict Amid Community: The Micropolitics of Teacher Collaboration. *Teachers College Record*, *104*(3), 421–455. https://doi.org/10.1111/1467-9620.00168

Akiba, M., & Wilkinson, B. (2016). Adopting an International Innovation for Teacher Professional Development: State and District Approaches to Lesson Study in Florida. *Journal of Teacher Education*, *67*(1), 74–93. https://doi.org/10.1177/0022487115593603

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Barab, S. A., & Luehmann, A. L. (2003). Building Sustainable Science Curriculum: Acknowledging and Accommodating Local Adaptation. *Science Education*, *87*(4), 454–467. https://doi.org/10.1002/sce.10083

Bond, N. (2014). *The power of teacher leaders: Their roles, influence, and impact*. Routledge.

Briscoe, C., & Peters, J. (1997). Teacher Collaboration across and within Schools: Supporting Individual Change in Elementary Science Teaching. *Science Education*, *81*(1), 51–65. https://doi.org/10.1002/(SICI)1098-237X(199701)81:1<51::AID-SCE3>3.0.CO;2-0

Bruce, C. D., Flynn, T. C., & Bennett, S. (2016). A focus on exploratory tasks in lesson study: The Canadian "Math for Young Children" project. *ZDM Mathematics Education*, *48*(4), 541–554. https://doi.org/10.1007/s11858-015-0747-7

Bruce, C. D., & Ross, J. A. (2008). A model for increasing reform implementation and teacher efficacy: Teacher peer coaching in grades 3 and 6 mathematics. *Canadian Journal of Education*, *3*(2), 346–370.

Bryk, A., Camburn, E., & Louis, K. S. (1999). Professional community in Chicago elementary schools: Facilitating factors and organizational consequences. *Educational Administration Quarterly*, *35*, 751–781. https://doi.org/10.1177/0013161X99355004

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

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.

Cravens, X., Drake, T., Goldring, E., & Schuermann, P. (2017). Teacher Peer Excellence Groups (TPEGs): Building Communities of Practice for Instructional Improvement. *Journal of Education Administration*.

Cravens, X., & Drake, T. (2017). From Shanghai to Tennessee: Developing instructional leadership through Teacher Peer Excellence Groups. *International Journal for Lesson and Learning Studies*, *6*(4), 348–364. https://doi.org/10.1108/IJLLS-12-2016-0062

Fujii, T. (2016). Designing and adapting tasks in lesson planning: a critical process of Lesson Study. *ZDM Mathematics Education*, *48*(4), 411–423. https://doi.org/10.1007/s11858-016-0770-3

Giles, C., & Hargreaves, A. (2006). The Sustainability of Innovative Schools as Learning Organizations and Professional Learning Communities During Standardized Reform. *Educational Administration Quarterly*, *42*(1), 124–156. https://doi.org/10.1177/0013161X05278189

Goddard, R. D., Hoy, W. K., & Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal*, *37*(2), 479–507.

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A Theoretical and Empirical Investigation of Teacher Collaboration for School Improvement and Student Achievement in Public Elementary Schools. *Teachers College Record*, *109*(4), 877–896.

Goldring, E., Porter, A. C., Murphy, J., Elliott, S. N., & Cravens, X. (2009). Assessing Learning-Centered Leadership: Connections to Research, Professional Standards, and Current Practices. *Leadership and Policy in Schools*, *8*, 1–36. https://doi.org/10.1080/15700760802014951

Groves, S., Doig, B., Vale, C., & Widjaja, W. (2016). Critical factors in the adaptation and implementation of Japanese Lesson Study in the Australian context. *ZDM Mathematics Education*, *48*(4), 501–512. https://doi.org/10.1007/s11858-016-0786-8

Hallinger, P., & Heck, R. H. (2010). Leadership for learning: Does collaborative leadership make a difference in school improvement? *Educational Management Administration & Leadership*, *38*(6), 654–678.

Hargreaves, A. (1994). *Changing teachers, changing times: Teachers' work and culture in the postmodern age*. London and New York: Continuum.

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher*, *31*(5), 3–15.

Huang, R., & Shimizu, Y. (2016). Improving teaching, developing teachers and teacher educators, and linking theory and practice through lesson study in mathematics: an international perspective. *ZDM Mathematics Education*, *48*(4), 393–409. https://doi.org/10.1007/s11858-016-0795-7

Jensen, B., Sonnemann, J., Roberts-Hull, K., & Hunter, A. (2016). *Beyond PD: Teacher Professional Learning in High-Performing Systems*. Washington, DC.

Leithwood, K., Seashore, K., Anderson, S., & Wahlstrom, K. (2004). Review of research: How leadership influences student learning.

*Lessons from Our Educators: Tennessee Educator Survey 2019 Results in Context*. (2019).

Lewis, C., Perry, R., & Murata, A. (2006). How Should Research Contribute to Instructional Improvement? The Case of Lesson Study. *Educational Researcher*, *35*(3), 3–14. https://doi.org/10.3102/0013189X035003003

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record*, *91*(4), 509–536.

Louis, K. S. (2006). Changing the culture of schools: Professional community, organizational learning, and trust. *Journal of School Leadership*, *16*(5), 477–489.

Louis, K. S., & Marks, H. M. (1998). Does professional community affect the classroom? Teachers' work and student experiences in restructuring schools. *American Journal of Education*, *106*(4), 532–575.

Louis, K. S., Marks, H. M., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal*, *33*(4), 757–798.

McDonald, L. (2012). Educational Transfer to Developing Countries: Policy and Skill Facilitation. *Procedia - Social and Behavioral Sciences*, *69*, 1817–1826. https://doi.org/10.1016/j.sbspro.2012.12.132

McLaughlin, M. W. (1987). Learning From Experience: Lessons From Policy Implementation. *Educational Evaluation and Policy Analysis*, *9*(2), 171–178. https://doi.org/10.3102/01623737009002171

Morris, A. K., & Hiebert, J. (2011). Creating shared instructional products: An alternative approach to improving teaching. *Educational Researcher*, *40*(1), 5–14.

OECD. (2011). *Successful Reformers in Education: Lessons from PISA for United States*. https://doi.org/http://dx.doi.org/10.1787/9789264096660-en

OECD. (2018). *Effective teacher policies: insights from PISA*. OECD Publishing.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data*. National Bureau of Economic Research.

Printy, S. M. (2008). Leadership for teacher learning: A community of practice perspective. *Educational Administration Quarterly*, *44*(2), 187–226. https://doi.org/10.1177/0013161X07312958

Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*, *44*(5), 635–674. https://doi.org/10.1177/0013161X08321509

Roehrig, G. H., Kruse, R. A., & Kern, A. (2007). Teacher and school characteristics and their influence on curriculum implementation. *Journal of Research in Science Teaching*, *44*(7), 883–907.

Rose, R. (1991). What Is Lesson-Drawing? *Journal of Public Policy*, *11*(1), 3–30.

SCORE. (2019). *Priorities for Progress: 2018-19 State of Education in Tennessee*.

Sindelar, P. T., Shearer, D. K., Yendol-Hoppey, D., & Liebert, T. W. (2006). The sustainability of inclusive school reform. *Exceptional Children*, *72*(3), 317–331. https://doi.org/10.1177/001440290607200304

Spillane, J. P. (1999). State and local government relations in the era of standards-based reform: Standards, state policy instruments, and local instructional policy making. *Educational Policy*, *13*(4), 546–572. https://doi.org/10.1177/0895904899013004004

State Report Card. (n.d.). Retrieved from https://reportcard.tnk12.gov/

Stigler, J. W., & Hiebert, J. (2016). Lesson study, improvement, and the importing of cultural routines. *ZDM Mathematics Education*, *48*(4), 581–587. https://doi.org/10.1007/s11858-016-0787-7

TDOE. (2016). *Teacher and Administrator Evaluation in Tennessee: A Report on Year 4 Implementation*.

TDOE. (2017). *Every Student Succeeds Act: Building on Success in Tennessee: ESSA State Plan*.

ten Bruggencate, G., Luyten, H., Scheerens, J., & Sleegers, P. (2012). Modeling the influence of school leaders on student achievement: how can school leaders make a difference? *Educational Administration Quarterly*, *48*(4), 699–732.

USDOE. (n.d.). International Data Explorer. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Wang, J. (2013). Introduction of school-based teacher professional development in China. *Asia Leadership Roundtable, Shanghai*.

Youngs, P., & King, M. B. (2002). Principal leadership for professional development to build school capacity. *Educational Administration Quarterly*, *38*(5), 643–670. https://doi.org/10.1177/0013161X02239642

Zech, L. K., Gause-Vega, C. L., Bray, M. H., Secules, T., & Goldman, S. R. (2000). Content-Based Collaborative Inquiry: A Professional Development Model for Sustaining Educational Reform. *Educational Psychologist*, *35*(3), 207–217. https://doi.org/10.1207/S15326985EP3503_6

## APPENDIX A: TEACHER INTERVIEW GUIDE

Note: Below are two types of interview guides, one for teachers and one for administrators. Because the status of TPEG varied by school and whether teachers had ever participated in TPEG varied by person, I have multiple versions of each guide, which I used with a skip pattern, per the bolded instructions at the beginning of each section. Not every question or section ended up being appropriate for the schools, so I updated it as needed. The guide was edited based on the amount of time afforded for each interview. When I had very little time, I focused on how their current collaborative planning works, how they feel about it, and what roadblocks they have encountered. I incorporated follow up questions frequently, as they seemed appropriate.

### Introduction

We talked about this over email, but as a reminder, my name is Olivia Carr, and I am a Ph.D. student at Vanderbilt University studying education policy. This study is about TPEG, the teacher collaboration model that started in your school a few years ago, as well as your views on collaboration and teacher learning in general.

- Do/did you and your colleagues call it TPEG when you talk(ed) together? If not, what do/did you call it?

Through this interview and interviews with your colleagues, I'm trying to understand what has made collaboration successful and challenging at your school. I'm hoping today to hear about your perspective on why you do or do not participate in TPEG and other forms of collaboration, supports and resources, and how your views and TPEG itself have changed over time. There are no right or wrong answers here; I'm just hoping to hear your honest thoughts and opinions. This interview is completely voluntary, and you can skip any questions you do not wish to answer. You may also end the interview early if you wish. Is there anything you want to know from me before we get started?

### Background & Style

1. Can you tell me about your teaching background and experience?
    a. How long have you been a teacher, both overall and at this school?
    b. What do you teach?
    c. How much experience did you have doing collaborative lesson planning before TPEG?
    d. How would you characterize those experiences?

    e. How were you trained to be a teacher? (traditional program, Teach for America, etc.)
    f. Did you do student teaching? Tell me about that experience.
2. Can you describe the nature of your participation in TPEG/collaborative planning?
    a. When/how many years did you participate in TPEG/collaborative planning?
    b. What teams have you been on?
    c. Have you ever been a TPEG leader?
3. How would you describe your usual teaching style?
4. What's your process to plan most of your lessons?
5. What does one of your typical lessons look like?
6. How do these things, mostly your style and the final lesson, change, if at all, when they are for TPEG lessons?
7. How do you judge whether a lesson has been successful?

### Early TPEG/Collaborative Planning

If they have been participating in TPEG/collaborative planning for a while…

1. I'd like to ask about your early days with TPEG/collaborative planning. How do you remember TPEG/collaborative planning being introduced to you and how you were trained to participate in it?
2. What was your reaction to it when you were first introduced to TPEG/collaborative planning?
3. Tell me about one of the first TPEG/collaborative planning cycles you participated in and when that was. Could you walk me through what happened?
    a. Who was there with you? What did they do?
    b. How did you choose what teacher, lesson, topic to focus on?
    c. How did your team choose when and how to meet? Face-to-face, online, video interactions, etc.
    d. What happened in the TPEG/collaborative planning cycle after the debrief and refinement of the lesson?
4. How would you have characterized TPEG/collaborative planning at that time? In other words what did you think of it?
    a. What did you think?
    b. What did others, such as other teachers and the administration, think?
    c. Why did you participate in TPEG/collaborative planning?

5. What was it that made TPEG/collaborative planning work or not work well for you and your team back then?
    a. What resources or challenges did you have inside and outside TPEG/collaborative planning meetings?
    b. *If necessary, suggest topics like logistics, resources, time, etc. to make sure they don't only think about interpersonal relationships and what happens inside the meetings.*

TPEG/Collaborative Planning Today

If they're still participating in TPEG/collaborative planning …

1. Let's fast-forward to today and talk about what, if anything, has changed with TPEG/collaborative planning since your earlier interactions with it. What does a TPEG/collaborative planning cycle look like now? Has the overarching structure changed much over time?
2. Walk me through a recent TPEG/collaborative planning cycle, perhaps the last one you participated in or another typical TPEG/collaborative planning cycle.
    a. *If necessary:*
    b. Who was there with you? What did they do?
    c. How did you choose what teacher, lesson, topic to focus on?
    d. How did your team choose when and how to meet? Face-to-face, online or video interactions, etc.
    e. Did anything happen after the debrief and refinement of the lesson?
    f. How have these particulars of a TPEG/collaborative planning cycle or how TPEG/collaborative planning works at your school changed since you first started participating, if at all?
    g. How prepared to you feel to carry out the work of TPEG/collaborative planning?
3. Nowadays, what do you think of TPEG/collaborative planning?
    a. Do you think that TPEG/collaborative planning is successful at your school? Why or why not? What are some of the challenges and strengths of TPEG/collaborative planning at your school?
    b. Do others think that TPEG/collaborative planning is successful? Why or why not?
    c. Why do you participate in TPEG/collaborative planning now?

        i. If it weren't mandatory/emphasized/_______ at your school, would you still try to participate?

4. What is it that makes TPEG/collaborative planning work or not work well now that it's been around for a while?
    c. What resources or challenges do you have inside and outside TPEG/collaborative planning meetings?
    d. *If necessary, suggest topics like logistics, resources, time, etc. to make sure they don't only think about interpersonal relationships and what happens inside the meetings.*
6. How do you know if a lesson has been successful?
5. If you were advising another school that was about to implement TPEG/collaborative planning, what would you tell them they need or should avoid to make TPEG/collaborative planning successful?
6. Do you know why other teams are not participating in TPEG/collaborative planning at your school?
7. What does a good TPEG/collaborative planning cycle look like? What does a bad TPEG/collaborative planning cycle look like?

If they stopped participating in TPEG/collaborative planning…

1. Let's fast-forward to more recently. Tell me about TPEG before you stopped participating in it. What did a TPEG cycle look like? Had the overarching structure changed much over time?
    a. What about the specifics, including who attended meetings, how you chose topics, how you met, things like that?
    b. How do you think that was different from your earlier times with TPEG?
    c. How prepared did you feel to carry out the work of TPEG?
2. Tell me about the events leading up to your team no longer participating in TPEG.
    a. What was it that made TPEG work or not work well those last times you participated?
        i. What resources or challenges did you have inside and outside of TPEG meetings?
3. What do you think would need to be required for TPEG to have been successful for your team?
    a. This can be contextual things, like a different group of students, personal things, like who's in the room, resources, like time, etc.
4. What do you do instead of TPEG during your PLC meetings?

5. What do you think would happen if your team were asked to try TPEG again?
6. If you were advising another school that was about to implement TPEG, what would you tell them they need or should avoid to make TPEG successful?

If they never participated in TPEG/collaboration…

1. Because you have not participated in TPEG, I am going to ask a few questions about your familiarity with and feelings about the model before moving on to other topics. Can you tell me what you know, if anything, about TPEG?
2. TPEG is a teacher collaboration model where a team of teachers come together to plan a lesson; then a few teachers teach the lesson while the others observe; the team then comes back together to provide feedback on the lesson and update it based on their observations; then the lesson is stored to be used in another year. What do you think of that model?
3. Would you want to participate in a model like that? Why or why not?
4. How do you think your current grade-level or departmental team would like participating in a model like that?
5. Do you know how others in your school, including other teachers and administration, feel about TPEG?

### Other Collaboration

1. Now that we've talked about TPEG, I'd like to hear about your thoughts on collaboration more generally. What are the ways, if any, that you have collaborated with your peers?
2. What would be your ideal in terms of collaboration with your peers? Why?
    a. *Note: Could be no collaboration*

### Professional Learning

1. Let's talk about professional learning, too. In ideal professional development or professional learning situation, what would you want to take away from it? What are you looking for in good professional development?
    a. This can be something like TPEG/collaborative planning, traditional PD sessions, or any other kind of professional learning.
2. Do you think what you're looking for has always been the same or has changed throughout the course of your years teaching?

    a. Do you think it will change as you continue teaching?

3. What about for observations? What do you hope to take away from the experience when you observe other teachers (or if you were to observe other teachers)?

Wrap Up

1. What do you prefer, your usual/old way of planning or TPEG/collaborative planning?
2. Is there anything else you think I should know about you, your school, or collaborative planning that you think would relate to these lines of questioning?

## APPENDIX B: ADMINISTRATOR/COACH INTERVIEW GUIDE

### Introduction

You might already know this, but as a reminder, my name is Olivia Carr, and I am a PhD student at Vanderbilt studying education policy. This study is about TPEG, the motivation for adopting it and participating in it, what resources and supports have been put in place, and how it's changed over the years it's been at your school. Is there anything you want to know from me before we get started?

### Background

1. Can you tell me about your background and experience in education?
    a. Were you a teacher before this?
        i. If so, how much experience did you have doing collaborative lesson planning?
        ii. How would you characterize those experiences? Was it something you liked?
2. How do you remember TPEG being introduced to you?
3. In your first impressions, what drew you to TPEG?
4. Were there any parts of it that made you hesitant to adopt it?
5. Why did your school start using TPEG?

### Early TPEG

1. Can you talk about making the initial TPEG teams?
2. How did TPEG go at first, before Shanghai? What went well, and what were some of the challenges?
3. If they went to Shanghai: I understand you were the one who went to Shanghai with Vanderbilt. Will you tell me about that experience?
    a. What were some of the major things you learned that you wanted to bring back home?
    b. How necessary do you think the Shanghai trip was to the success of TPEG?
4. What do you think changed in how TPEG worked here after the Shanghai trip?
5. Will you tell me how TPEG worked in that first year?
    a. How long were the cycles?
    b. How were meeting and observation times scheduled?
    c. Were they always in person or were there discussions of doing it over video or another way?

    d. Who went to the meetings?
    e. How did the teams choose what teacher, lesson, topic to focus on?
6. How would you have characterized TPEG at that time? What did you think of it?
    a. What did others, such as the teachers and others in administration think?
    b. Why did the teachers participate in TPEG?
7. What was it that made TPEG work or not work well in the early days?

If TPEG Expanded…

1. What was going on behind the scenes for the expansion of collaborative planning? What decisions were being made? What conversations were being had?
2. What did the teachers think of this expansion, both the ones who had participated before and those who hadn't?
3. Besides, of course, the number of teachers and teams participating, was there anything else you decided to change about how TPEG worked when you expanded it?
4. Let's fast-forward to today. What are the major things that have changed with TPEG/collaborative planning over time?
    a. Who goes to those meetings? Do administrators or coaches join?
    b. Do you have a particular way to store the collaborative planning lessons?
5. What are the major challenges with collaborative planning?
    a. At other schools, I've heard that some teachers feel like they're putting on a show, not being authentic, when planning TPEG/collaborative planning lessons, and the final product is very different, in a bad way, between their normal lessons and a TPEG/collaborative planning lesson. Has that been a problem?
    b. I also know that some people have a hard time being observed or being a "critical friend." Do you think that's been a problem here?
6. What are the major parts that have been successful?
7. Have there been breaks in using TPEG/collaborative planning over the years or has it been fairly consistent during the school year since you first started using it?
8. Do you have ways to document changes from TPEG/collaborative planning, maybe in teacher working conditions, quality of lessons, student achievement?
9. What do you think of collaborative planning now?
    a. Do you think that collaborative planning is "successful" at your school?

   b. What do others, like teachers and other administrators, coaches, think of collaborative planning?
   c. Has collaborative planning been equally successful for all teams?
   d. Do parents have any kind of role with collaborative planning, maybe with coverage? Do they know anything about it?
   e. What would happen if you had teachers who said they no longer wanted to participate in collaborative planning? Either teachers who have participated for a while or new hires.

10. What does a good collaborative planning session look like? What does a bad collaborative planning session cycle look like?
11. If you could change something about collaborative planning or your experience with TPEG, what would it be?
12. If you were advising another school that was about to implement TPEG/collaborative planning, what would you tell them they need or should avoid to make TPEG/collaborative planning successful?

## Collaborative Planning Moving Forward

If TPEG Stopped…

1. Let's fast-forward. How long did TPEG last?
2. Before it was let go, had you adapted TPEG at all from what it looked like when was introduced by Vanderbilt?
3. What were the major challenges with TPEG?
   a. At other schools, I've heard that some teachers feel like they're putting on a show, not being authentic, when planning TPEG lessons, and the final product is very different, in a bad way, between their normal lessons and a TPEG lesson. Was that a problem here?
   b. I also know that some people have a hard time being observed or being a "critical friend." Do you think that was a problem here?
4. What were the major parts that were successful?
5. Why did TPEG end at your school?
6. What do you think would need to be required for TPEG to have been successful for your school?
   a. This can be contextual things, like a different group of students, personal things, like who's in the room, resources, like time, etc.

7. Would you consider trying TPEG again? Why or why not?
8. What collaboration is in place now, if any?
9. Would you consider a different type of collaborative planning?
    a. How would you encourage teachers to use that?
10. If you could change something about TPEG or your experience with TPEG, what would it be?
11. If someone from another school were thinking about implementing TPEG, what would you say?

## Wrap Up

1. Is there anything else you think I should know about you, your school, collaboration, or professional learning that you think would relate to this line of questioning?

# CHAPTER 3

## THE USE OF STUDENT ASSESSMENTS TO EVALUATE TEACHERS: AN INTERNATIONAL, LONGITUDINAL STUDY

For over half a century, nations around the world have been increasingly interested in conducting national learning assessments (NLAs) of their students (Best et al., 2013; Bruns, Filmer, & Patrinos, 2011; Kamens & McNeely, 2010; Postlewaite & Kellaghan, 2008). Some of this rise came from intergovernmental organizations as a condition of loans in order to “rationalize” education systems in the developing world, but much of it came from an international ideological shift toward globalization and the belief that countries need to have national and/or international testing to be successful as a society (Kamens & McNeely, 2010). The World Conference on Education for All in Jomtien, Thailand, in 1990 promoted the increasing use of NLAs by emphasizing that access to schools is only impactful if students are actually learning within them (Postlewaite & Kellaghan, 2008), and countries were more frequently considering student test scores as the central measure of educational quality (Bruns et al., 2011). Since then, the number of countries with NLAs has been increasing rapidly, from 65 countries having at least one NLA in 1995-1999 to 111 countries having an NLA in 2000-2006. Developed countries are more likely to have NLAs, but developing countries are increasingly using them as well; 81 percent of developed countries and 51 percent of developing countries administered an NLA as of 2006 (Benavot & Tanner, 2007). The Leadership Council of the Sustainable Development Solutions Network (2015) reports that more than 60 developing countries have used national or international comparable reading assessments since 2005. The goals of these NLAs are to evaluate the quality and equity of the nation’s schools and, sometimes, to hold individual schools and teachers accountable for their students’ learning (Best et al., 2013).

There have been higher levels of monitoring and accountability in education at every level as schools are increasingly being painted as bureaucratic organizations that should be efficient, high functioning, and objectively evaluated in a standardized manner (Hanberger, 2013; Kamens & McNeely, 2010; Lindgren, Hanberger, & Lundström, 2016; Meyer & Benavot, 2013; Smith, 2016; Tatto, 2006). The theory of new institutionalism describes this process as tighter coupling between schools and other organizations that seek to control and monitor educational practices at an increased level (Meyer & Rowan, 2006). Teacher evaluation, in particular, serves various purposes, including holding teachers to a high standard of teaching and improving the quality of teaching through professional learning (Dahler-

Larsen, 2012; Danielson, 2001; Isoré, 2009; Tatto, 2006). As students' learning is being measured by tests, teacher evaluators are sometimes including student test scores in these evaluations as a supposedly objective measure of the teachers' teaching abilities (Isoré, 2009). The goal of including student test scores in teacher evaluations is to hire and fire teachers based on their effects on student learning and to motivate teachers to become better and to put student learning at the center of their teaching (Adnot, Dee, Katz, & Wyckoff, 2017).

My study seeks to test, in Organisation for Economic Co-operation and Development (OECD) and partner countries, the notion that comes out of new institutionalism that including student test scores in teacher evaluations will improve student learning. In particular, I aim to determine to what extent the inclusion of student test scores in teacher evaluations changes student academic achievement across 59 OECD and partner countries. I evaluate data from four waves of the Programme for International Student Assessment (PISA) to address this question. The PISA test is an appropriate indicator of  student learning  because it does not have a direct effect on students' or teachers' lives, which should minimize the influence of some of the more nefarious side effects of high-stakes assessments, like teachers changing students' incorrect answers to correct ones (Hanushek & Rivkin, 2010). The key methodological component of this analysis is the use of a country-time aggregate measure of whether schools use student test scores to evaluate teachers; my approach controls for selection bias that would arise at the school level. To my knowledge, questions around the use of student assessment data in teacher evaluations have not been addressed quantitatively at the international level beyond basic descriptive statistics.

## Literature Review

The education production function literature posits that a student's learning is comprised of a variety of inputs, including factors that are outside schools, like family social and cultural status, and those that are within schools, like the quality of the student's teachers (for example, Hanushek, 1986). Teacher evaluation is intended to improve the overall quality of teaching (Bruns et al., 2011; Darling-Hammond, 2004; Isoré, 2009; Leithwood & Earl, 2000; Wößmann, 2007).

Many different sources can be used to evaluate teachers. Evidence of teacher performance can come from student outcomes, tests given to the teacher, surveys given to students or parents, classroom observations, interviews with the teacher, and documentation prepared by the teacher; researchers and practitioners often believe that multiple sources should be used to increase the accuracy and fairness of the evaluation (Peterson, 2000). The most common source of evidence comes from classroom observations, by an external observer, peer teacher, or administrator within the school, though the results from observations are not always informative of teacher quality (Isoré, 2009). Using student test scores in evaluations is also becoming common practice in countries that participate in the OECD Teaching and Learning International Survey (TALIS; Smith & Kubacka, 2017).

The theory of new institutionalism predicts that institutions will undertake policies and activities that promote their legitimacy as institutions and more tightly couple institutional goals and practices. These concepts have led to an increased demand for productivity and accountability in education (Meyer & Rowan, 2006). In an effort to use objective and fair measures of teacher quality, some education systems place an emphasis on student learning outcomes and, in particular, student test scores, when they evaluate teachers. The desire to use numbers and standardized metrics in evaluation reflects the global accountability and audit culture that has developed (Dahler-Larsen, 2012; Hardy & Boyle, 2011), and government officials are becoming increasingly convinced that student test scores are the key measure of a successful education system (Bruns et al., 2011). The use of student learning outcomes in teacher evaluation demonstrates that the education system prioritizes student performance (Anderson, 2005) and emphasizes that improvement in student learning should be the primary goal of teaching (Isoré, 2009).

Evaluators can consider student test scores in teacher evaluations in a variety of ways. Test data can be officially incorporated into a standardized evaluation system, such as with the District of Columbia (D.C., United States) IMPACT evaluation system that was implemented in 2009-10; student value-added assessment data comprised 50 percent of qualifying teachers' evaluation scores in this case (Adnot et al., 2017). In other systems, student test score data can be incorporated into teacher evaluations at the discretion of the evaluator. For instance, in Austria, some teachers receive their students' assessment data and may choose to incorporate this information in their self-evaluations (Specht & Sobanski, 2012). The type of test considered can also vary, from a teacher-created test for her own students to a countrywide standardized assessment that every student is required to complete. Some systems evaluate teachers or schools on the same standards of test scores, despite differences in resources or the backgrounds of the students, but researchers tend to prefer evaluation policies that take into account the initial achievement levels of the student prior to exposure to the teacher, or value-added (Darling-Hammond, 2004; Isoré, 2009).

Measures that are frequently used for decisions on hiring and compensation, like years of experience or education level, often do not explain much of the variation in teacher quality, so value-added and other systems that incorporate student assessment data  aim to provide valuable information to improve decisions regarding teachers (Hanushek & Rivkin, 2010). Researchers often conclude that evaluation based on student assessment data should not be the only type of evidence included in teacher evaluations but can be useful in an evaluation system, and evaluation systems internationally tend to include multiple measures in their teacher evaluation systems (Smith & Kubacka, 2017).

The goal of evaluating teachers and, in particular, evaluating teachers using student assessment data is to improve student learning. The theory of change, shown in Figure 1, has multiple and often simultaneous paths by which this might happen, including hiring high-performing teachers, firing low-performing teachers, encouraging improvements in the quality of teaching among existing teachers, and providing other incentives for teachers to work hard and to put student learning at the center of their craft (Bruns et al., 2011; Danielson, 2001; Darling-Hammond, 2004; Isoré, 2009; Leithwood & Earl, 2000; Wößmann, 2007). These changes lead to improvements in teacher quality, and there is a strong literature that links teacher quality with student achievement (Akiba, LeTendre, & Scribner, 2007; Darling-Hammond, 2000).

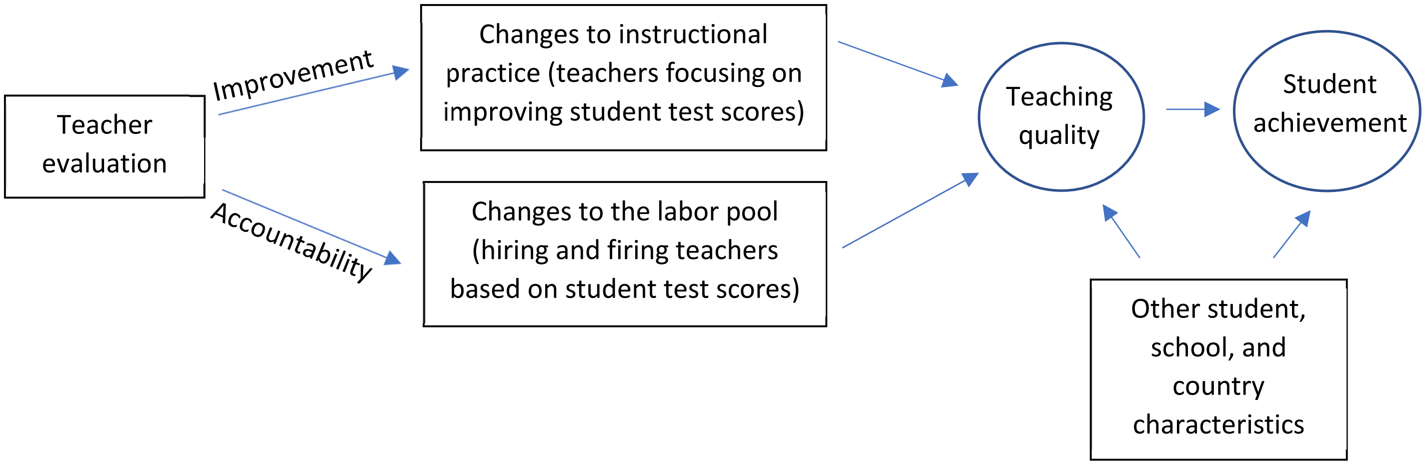

Teacher evaluation
Improvement
Accountability
Changes to instructional practice (teachers focusing on improving student test scores)
Changes to the labor pool (hiring and firing teachers based on student test scores)
Teaching quality
Student achievement
Other student, school, and country characteristics

Figure 1: Teacher Evaluation Theory of Change

The controversial D.C. IMPACT teacher evaluation system uses a variety of sources for teacher evaluations, including student value-added data and observations, for high-stakes decisions regarding teacher dismissals as well as raises and bonuses. IMPACT also provides feedback on evaluations to teachers and support for teachers to improve their practice (Adnot et al., 2017; IMPACT, 2018). Adnot, Dee, Katz, and Wyckoff (2017) reviewed teacher turnover in this system in the first four years of its existence and found that, under IMPACT, low-performing teachers exited the school system at three times the rate of high-performing teachers. They also found that IMPACT likely improved student achievement, particularly in high-poverty schools, where the exit of low-performing teachers increased student math performance by 20 percent of a standard deviation and reading performance by 14 percent of a standard deviation.

From 2006 until a new system was implemented in 2015, Mexico had a student assessment regime that was high-stakes for teachers and schools. The test was called the National Assessment of Academic Achievement in Schools, or ENLACE. Part of teacher and administrator salaries were tied to raw ENLACE scores, meaning there were no corrections for student or school characteristics, including prior student achievement (OECD, 2018; Santiago, Gilmore, Nusche, Ravela, & Sammons, 2012). Some teachers reported to the OECD that the system applied positive pressure on teachers to focus on improving student achievement, but many others thought it to be an affront to their professionalism (Santiago, Gilmore, et al., 2012). In Mexico's new teacher appraisal system, implemented in 2015-16, teachers get individualized coaching the first or second time they fail the evaluation and get fired the third time. The new student assessment, called the National Plan for Learning Assessment, or PLANEA, is primarily used as a formative assessment and does not have formal consequences for teachers or schools (OECD, 2018).

Portugal implemented a new teacher pay scale system in 2007 that used teacher performance to move teachers along the scale. Criteria for the pay increase included the teacher's attendance, involvement in research, and, controversially, the incorporation of student assessment data. The goals of the evaluation system were to improve teacher practices, reward excellence, and provide data to management about the performance of teachers (Santiago, Donaldson, Looney, & Nusche, 2012). Using a difference-in-differences approach, Martins (2009) found that this teacher pay scale incentive system resulted in lower student achievement on national exams and increased grade inflation. The 2007 teacher evaluation system faced so much resistance that it was updated after a few years to move away from using student outcomes to focusing on other dimensions of teaching (Santiago, Donaldson, et al., 2012).

## Data & Methods

This study seeks to identify the effects of using student assessment data to evaluate teachers on student academic performance. There are several problems that would bias the results of a model that simply regressed the academic performance of students in a school on the existence of this method of evaluation in that school, including selection bias. Following the methodology used in Hanushek, Link, & Woessmann (2013), I account for these problems with the aggregation of the key variable of interest to the country-year level and the inclusion of a series of fixed effects and country-level control variables. The student assessment data need not come from a national learning assessment and can thus vary by school. The most serious concern in this study is selection bias, with families and/or teachers choosing schools based on the presence or lack of a teacher evaluation system that includes student assessment data. Another concern with selection bias is whether only certain types of schools, perhaps those with students who test well, adopt teacher evaluation models that incorporate student assessment data. With the assumption that these concerns would never or rarely be an issue across country borders, I remove this selection bias by aggregating the presence of this evaluation system to the country level for each wave of data. The key independent variable of interest is the proportion of students who attend a school that uses student assessment data to evaluate teachers in a country-year.

To remove time-invariant factors within countries, I include country fixed effects. Year fixed effects control for common shocks across waves, including updates to the norming of the tests. The final concern is that there could be time-varying country factors that are associated with the presence of using student data to evaluate teachers that would also relate to trends in student achievement. Most likely, one time-varying country factor would be the presence of other educational accountability systems that are introduced at a similar time as using student data to evaluate teachers. For example, the introduction of a new teacher observation protocol could be incorporated into teacher evaluations at the same time as the use of student assessment data in teacher evaluations is introduced. The underlying assumption is that there are no such time-varying country factors, but I include a series of time-varying country variables in some models to help assuage concerns with this issue. One of those variables specifically seeks to capture other educational accountability systems that might be present.

The final model is as follows:

$$Y_{ctis} = B_0 + B_1{*}TE_{ct} + B_S{*}S_{ctis} + B_I{*}I_{cti} + B_C{*}C_{ct} + u_c + u_t + v_{ctis}$$

Y is the student outcome, either math or reading test scores, at the country-time-institution (school)-student level. The key independent variable of interest is TE, the proportion of schools in a given year and country that use student assessment data in teacher evaluations. I control for a series of student and institution characteristics to improve the precision of the estimates, but they should not affect the direction and magnitude. Time-varying country characteristics attempt to account for the relationship between TE and the error term *v*. Finally, the model includes country and time fixed effects.[5]

*Data*

The data primarily come from the Programme for International Student Assessment (PISA), and the GDP per capita and educational expenditures variables come from the World Bank Open Data portal. PISA tests samples of students around the world once every three years, starting in 2000. Four of the first five waves of PISA data have PISA math and reading performance data and several longitudinally-consistent questions posed to the principal about various uses for student assessment data within the school. Table 1 displays the weighted percent of students who are in a school in which the principal answered affirmatively to the survey questions. It shows very high rates of schools using student assessment data to inform parents of their child's progress in school and lower, but perhaps increasing, rates of using it for grouping students. The use of student assessment data for student accountability, in particular retention or promotion from one grade to the next, is slightly decreasing over time. The last two questions about school and teacher accountability and monitoring are increasing across the study years, and they contribute to this particular study.

Table 1: Percent of Students Who Attend a School that Uses Student Assessment Data for Actions

| **Survey Question** | **2000** | **2003** | **2009** | **2012** |
|---|---|---|---|---|
| *In your school, are assessments of students in the 10th grade used for any of the following purposes?* | | | | |
| To inform parents about their child's progress | 98% | 95% | 98% | 98% |
| To make decisions about students' retention or promotion | 86% | 83% | 85% | 80% |
| To group students for instructional purposes | 55% | 54% | 63% | 62% |
| To monitor the school's progress from year to year | 80% | 80% | 89% | 90% |
| To make judgments about teachers' effectiveness | 64% | 64% | 70% | 72% |

Note: Survey weights incorporated.

The population of interest for PISA is all 15-year-olds in OECD and partner countries who are in an educational or vocational institution. Countries decide to participate in PISA for a variety of reasons, including a drive to best understand evidence-based policies, the promotion of accountability and transparency, pressures to follow international norms, and for global prestige (Addey & Sellar, 2018). For almost every country, PISA uses a two-stage, stratified, probability proportional to size sampling

[5] Analyses were conducted using Stata 14.

method, with schools in the first stage and students in the second. In particular, schools that are educational or vocational institutions that serve students who are 15 years old in each country were stratified (by variables such as region, funding type, or urbanicity) and then systematically sampled. A group of approximately 35 students was selected from each school with equal probability of being chosen. Exclusions of schools or students were minimized but were sometimes necessary, so occasionally remote geographical regions or students with special needs who could not provide responses or with very limited proficiency in the test language were excluded. PISA requires that the exclusion rate of schools and students be below 5 percent, and there are other exclusion requirements listed in the technical reports.[6] Selected students were given mathematics, reading, and science assessments – except in 2000, when students were randomly assigned one, two, or three of the subjects[7] – as well as a contextual questionnaire about their life and learning experiences. Their principals were asked to fill out a contextual questionnaire about the school as well. The minimum response rate for schools was 85 percent, and replacement schools were added if that response rate was not met. According to the technical reports, the highest overall exclusion rates tended to be in high-income countries, such as the US, Denmark, and Luxembourg.

There are a variety of potential limitations with PISA. Mortimore (2009) lists a variety of potential concerns with PISA, including the effects of the league (ranking) tables and narrow "economic" interests (testing, evaluation, efficiency, competitiveness) on education policy. While these are valid concerns, PISA does take steps to ameliorate major issues. For instance, with sampling, the result reports mark or move to the appendix countries that have abnormal samples for reasons such as high rates of replacement schools. The technical reports go into further detail about quality control, including how they hire and disperse independent PISA Quality Monitors to observe test administration in each country and ensure the testing experience across schools and countries is as identical as possible. PISA has introduced additional optional subjects that many education systems prioritize: financial literacy, problem solving, and collaborative problem solving. I do not include analyses with these tests because test scores are typically only included in teacher evaluations for teachers of major subjects (math, reading, and sometimes science and/or social studies), and the two problem solving tests are cross-sectional. While PISA is limited because it prioritizes certain subjects, the theory I am testing directly addresses a trend related to learning in primarily math and reading, so my use of only these subjects is appropriate.

[6] The technical reports for each wave of PISA can be found on the Data page of the PISA website.
[7] For the 2000 wave of PISA, I only include the students who took both the math and reading assessments.

The primary analytic sample for this study consists of 1,304,482 students in 51,996 schools in 59 countries across four waves of data, and I apply sampling weights to make the sample representative of all students in the included country-years. I excluded countries if they did not participate in PISA for at least two of the four waves of data or if they had very high levels of systematic missingness; for instance, France was excluded because it did not administer the school survey in two of the waves.

*Variables*

The outcome of interest is student achievement. Students take math, reading, and science tests for PISA, but I focus on math and reading. Rather than having a single math score and single reading score, PISA reports five plausible values for each subject, which are created using an imputation methodology to account for uncertainty inherent in estimating a students' academic knowledge. I account for these plausible values in the analysis by using MacDonald's (2008) plausible values command in Stata. One usual concern with using student assessment data to evaluate teachers is that teachers could manipulate assessment scores, perhaps by teaching to the test or changing student test answers. If manipulation is a significant problem the resulting test scores would not accurately reflect the learning of the student. However, the PISA tests are low-stakes assessments, at least for the students who take them and their teachers. Teachers tend to not teach specifically to the PISA tests, and there are no specific rewards or sanctions for students, teachers, or schools attached to the results. Therefore, these assessment results should more accurately reflect students' knowledge than the other, high-stakes assessments that might be used in teacher evaluations.

The key variable of interest is the proportion of students in a country-year who are in a school that uses student assessment data to evaluate teachers. I constructed this variable using answers from the school surveys. In the 2000, 2003, 2009, and 2012 waves, principals were asked, "In your school, are assessments of students in the 10th grade used for any of the following purposes?" They were given several situations, including "To make judgments about teachers' effectiveness." The principals were asked to answer yes or no for each. While this question does not explicitly reference teacher evaluation, and certainly not high-stakes teacher evaluation, I find that teacher evaluation policy changes predict changes in the responses to this question for several included countries, as shown later in this section. I therefore consider this question to represent how principals filter policy and other messages in making "judgments" or what I call "evaluations" about his/her teachers. The 2006 and 2015 questionnaires included similar, but not identical, questions. See Appendix A for more information on the 2006 wave, including how the analyses differ when including 2006 data. The 2015 question is different enough from that used in my analysis that I exclude it from this study.

To construct the key variable, I took the weighted percent of students in each country in each year who attend a school that uses student assessment data on teacher evaluations. According to new institutionalism, countries should feel pressure to use this tightly-coupled accountability system to maintain legitimacy as an institution, which predicts that this type of evaluation system will be more frequently used over time. Figure 2 shows this pattern for the whole sample and each included region. The percent of students in a school with this type of evaluation system is increasing on average over time. The average is highest in Asia, where this evaluation system is most prevalent, and lowest in Australia, New Zealand, and Tunisia, the three countries in the "Other" category. Europe is the only region that does not consistently increase over the time span.

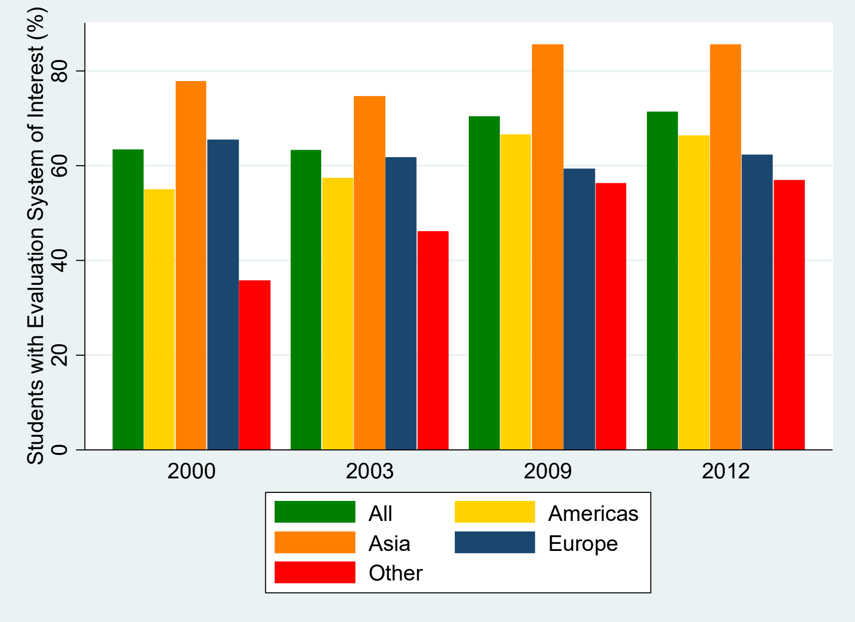


Figure 2: Percent of Students Who Attend a School that Uses Student Assessment Data When Evaluating Teachers by Region

Table 2 details the percentages by country and year. It shows that the lowest percent is Luxembourg in 2000, when zero percent of students were in a school that evaluated teachers using student assessment data. The highest percent is in Kazakhstan in 2012 with 100 percent of students attending a school that evaluated teachers using student assessment data. These extreme values can be explained by looking deeper at the two distinct contexts. Teachers in Luxembourg are rarely evaluated after their two-year probationary period, and only in exceptional situations is the evaluation used to fire a teacher. There was a federal push in the 2000s to strengthen the assessment and evaluation system, but some educators felt that evaluation was threatening and did not improve teaching quality (Shewbridge, Ehren, Santiago, & Tamassia, 2012). Unlike Luxembourg, the teacher evaluation system in Kazakhstan overemphasizes accountability. Teachers have to be recertified every five years, and one component of their application is information on the educational achievement of their students. This information is generally raw assessment scores and other student achievement measures (Pons et al., 2015).

My analysis includes country fixed effects, so Table 2 also shows which countries are likely to drive the results. The countries with the largest percentage point decrease over time are Italy, Finland, and Poland. The likely explanation for the rapid decrease in the incorporation of student assessment data in teacher evaluations in Italy from 2000 to 2003 is a 1997 law and 1999 presidential decree that increased school autonomy as of the 2000-2001 school year (Norlund, Marzano, & De Angelis, 2016).

Table 2: Percent of Students Who Attend a School that Uses Student Assessment Data When Evaluating Teachers

| Country Name | 2000 | 2003 | 2009 | 2012 |
|---|---|---|---|---|
| Albania | 91% | | 90% | 87% |
| Argentina | 64% | | 53% | 51% |
| Australia | 33% | 34% | 44% | 50% |
| Austria | 19% | 36% | 26% | 39% |
| Belgium | 26% | 19% | 31% | 35% |
| Brazil | 46% | 56% | 80% | 80% |
| Bulgaria | 89% | | 92% | 93% |
| Canada | 32% | 31% | 35% | 30% |
| Chile | 55% | | 58% | 61% |
| Chinese Taipei | | | 58% | 48% |
| Colombia | | | 63% | 60% |
| Costa Rica | | | 57% | 71% |
| Croatia | | | 55% | 56% |
| Czech Republic | 67% | 62% | 60% | 63% |
| Denmark | 4% | 4% | 8% | 27% |
| Estonia | | | 72% | 65% |
| Finland | 38% | 32% | 24% | 16% |
| Germany | 12% | 12% | 22% | 24% |
| Greece | 18% | 15% | 22% | 14% |
| Hong Kong | 57% | 64% | 76% | 80% |
| Hungary | 69% | 77% | 60% | 58% |
| Iceland | 43% | 31% | 29% | 39% |
| Indonesia | 97% | 87% | 98% | 96% |

| | | | | |
|---|---|---|---|---|
| Ireland | 29% | 17% | 37% | 47% |
| Israel | 82% | | 85% | 82% |
| Italy | 86% | 23% | 20% | 30% |
| Japan | 81% | 82% | 78% | 76% |
| Jordan | | | 82% | 72% |
| Kazakhstan | | | 99% | 100% |
| Latvia | 86% | 86% | 92% | 93% |
| Lithuania | | | 71% | 74% |
| Luxembourg | 0% | 21% | 22% | 22% |
| Macao (China) | | 82% | 74% | 75% |
| Malaysia | | | 92% | 92% |
| Mexico | 77% | 77% | 80% | 77% |
| Montenegro | | | 56% | 92% |
| Netherlands | 47% | 42% | 50% | 68% |
| New Zealand | 48% | 53% | 61% | 68% |
| Norway | 10% | 19% | 24% | 30% |
| Peru | 78% | | 80% | 78% |
| Poland | 93% | 73% | 79% | 79% |
| Portugal | 24% | 35% | 35% | 50% |
| Qatar | | | 84% | 87% |
| Romania | 20% | | 85% | 75% |
| Russia | 99% | 99% | 98% | 99% |
| Singapore | | | 85% | 88% |
| Slovak Republic | | 75% | 80% | 69% |
| Slovenia | | | 40% | 38% |
| South Korea | 14% | 54% | 66% | 85% |
| Spain | 42% | 36% | 44% | 50% |
| Sweden | 12% | 21% | 22% | 44% |
| Switzerland | 6% | 37% | 41% | 36% |
| Thailand | 76% | 71% | 95% | 91% |
| Tunisia | | 63% | 76% | 67% |
| Turkey | | 34% | 71% | 71% |
| United Arab Emirates | | | 91% | 94% |
| United Kingdom | 78% | 86% | 83% | 88% |
| United States | 56% | 55% | 58% | 60% |
| Uruguay | | 41% | 37% | 31% |

Note: Survey weights incorporated.

The countries with the largest percentage point increase over time are South Korea and Romania, though there are large increases in countries from most regions, including Asia (South Korea, Hong Kong), South America (Brazil), and Europe (Romania, Turkey, Montenegro, Sweden, Switzerland, Netherlands, Portugal, Denmark). South Korea has had performance evaluations conducted through self-evaluations, administrators, and peer teachers since the 1960s. In 2001, a performance-based pay system was developed and then updated in 2008 and 2010 to include increased school autonomy and a stronger professional development component. With some guidance from the Ministry of Education, schools have autonomy over what to include in the performance-based pay evaluations, and some schools decided to use student test scores in their teacher evaluations (Choi & Park, 2016; Sung Tae Jang, 2016), which is reflected in the increase in their reported use over time. Romania attempted to increase transparency, professionalism, and accountability over the study period. Teachers are required by law to participate in a substantial number of evaluations over the course of their careers, and the evaluations include student assessment data. The 2005 Quality Assurance Law and 2011 Education Law introduced new exams for students and accountability systems for teachers, which is reflected in the increased percentages over time for Romania in Table 2.

The control variables at the student level are gender, an indicator variable that is 0 for male and 1 for female, and economic, social, and cultural status (ESCS). ESCS is similar to the traditionally-used construct of socioeconomic status and controls for the student's family background. It is comprised of information on parental occupation, parental education, and home possessions. PISA created a rescaled index of ESCS with the 2015 data release to be used in trend analyses; I use the rescaled index of ESCS in my analysis. The index is continuous and, for my sample, ranges from -7.5 to 5.1. The control variables at the school level are size of the town, from large city (coded as 1) to a small city (coded as 0), whether it is a private school (1) or public school (0), whether it considers prior academic achievement in deciding who to admit (1) or does not use this information (0), and the number of students.

At the country level, I am primarily concerned with anything that would affect both students' test scores and the likelihood that the country has the teacher evaluation system that incorporates student assessment data. Most of those factors would likely be reflected in the funding directed toward education, for which I control using GDP per capita based on the purchasing power parity (PPP, in constant 2011 international dollars) and government expenditures on education as a percent of the GDP. I log transform GDP per capita to correct the skew in the variable. Countries might also refocus their existing educational funding toward other types of accountability and monitoring. The four waves of school surveys have a question that asks, "In your school, are assessments of 15-year-old students used to: monitor the school's progress from year to year?" with a yes/no response option. I use answers to this question to create a variable that I call *Accountability.* It controls for time-varying country-level changes in the accountability system that are separate from the inclusion of student assessment data in the teacher evaluation system within countries over time.

Table 3 shows the descriptive statistics for each variable, pooled across countries and years. It shows that the test scores have a mean of over 450 and a standard deviation of around 100, and the proportion of students who are in schools with this type of evaluation system is 68 percent. Appendix B shows descriptive statistics for each variable by country and pooled across years.

Table 3: Descriptive Statistics on Pooled Sample

| | Weighted Mean/Percent | Sample Standard Deviation | Sample N |
|---|---|---|---|
| *The average student has a(n)…* | | | |
| **Math Score** | **451.35** | **104.85** | **1,304,482** |
| **Reading Score** | **458.79** | **102.96** | **1,304,482** |
| ESCS | -0.734 | 1.20 | 1,279,158 |
| *X% of students are…* | | | |
| Female | 50.4% | - | 1,302,782 |
| *X% of students go to a school that is/has…* | | | |
| In a Large City | 40.7% | - | 1,254,692 |
| Private | 18.2% | - | 1,246,421 |
| Admission Based on Prior Academics | 57.3% | - | 1,259,835 |
| Monitoring School's Progress (*Accountability*) | 85.0% | - | 1,304,482 |
| **Evaluating Teachers using Student Data** | **67.8%** | - | **1,259,407** |
| *The average student goes to a school that has…* | | | |
| Number of Students | 763.75 | 613.27 | 1,217,856 |
| *The average student is in a country with…* | | | |
| GDP per Capita Based on Purchasing Power Parity (constant 2011 international $) | $25,971.92 | $19,820.70 | 1,292,605 |
| GDP per Capita (log transformed) | 9.98 | 0.59 | 1,292,605 |
| Education Expenditures as a % of GDP | 4.45% | 1.12% | 997,689 |

Note: Survey weights incorporated in the mean/percent. Bolded values are the dependent variables and the key variable of interest.

### *Missing Data*

Values are missing for 12% of the observations in models without country covariates and 33% of the observations in models with country covariates. The outcome variables, student math and reading scores, are not missing for any students. Prior to data cleaning, six education systems had high rates or complete missingness for important covariates, so I dropped those education systems from the analysis. France had high rates of missingness for the key predictor, and Azerbaijan, Kyrgyzstan, Lichtenstein, Shanghai (China), and Serbia were entirely missing ESCS. In the updated sample, the key variable is missing in no more than 6 percent of observations in each wave of data, and I detect no patterns of missingness for it.

For the student and school level variables, I was able to detect some patterns of missingness that were represented by differences in average test scores. I include test scores in the imputation models to help account for these differences.[8] There were also some patterns of missingness by year and by country. For instance, 40 percent of students in Albania do not have a value for the number of students in their school. Number of students is the only variable for which missingness does not seem to be well-explained by other variables in the model.

GDP per capita and education expenditures as a percent of GDP are missing for some countries, as can be seen in Table B3 in the appendix, and some country-years, based on the availability of World Bank data. I do not impute missing values for the country level variables, but the estimates for all of the covariates are relatively stable across models that have various covariates included and use various methods of dealing with missing data. As such, the missing data on the country covariates do not appear to significantly bias the results.

I only run the analyses on cases for which the key predictor and the country level variables are not missing. For the student and school level covariates, I impute missing values.[9] Due to theoretical and programming limitations, there are not many differences between using single and multiple imputation for the models I include in this paper. Single imputation does not account for the uncertainty inherent to the imputed data, and multiple imputation does not account for the uncertainty of the plausible value test scores. The estimates are qualitatively similar when using both methods, but the single imputation appears to account for more of the uncertainty, as per the size of the standard errors, so I use single imputation to account for missing data.

The variables with imputed values are gender, ESCS, whether the school is in a large city, whether it is private, number of students, and whether the school uses academics in admission. I use each of those variables and student math and reading scores in the imputation model.

## Results

The results in Tables 4 and 5 show the math and readings models, respectively. Model 1 includes year fixed effects to control for variations in the scoring of the test over time but does not include any country fixed effects or covariates. The estimate for Model 1 in Table 4 is negative and statistically significant, which indicates that country-years with lower average student achievement are the ones that have a higher prevalence of teacher evaluation that incorporates student test scores.

[8] The results are qualitatively similar when excluding student test scores from the imputation model.

[9] I employ single imputation using chained equations with Stata’s suite of `mi` commands.

Table 4: Average Student Achievement Scores on a Mathematics Test, Results on Pooled Sample

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Teacher Evaluation using Student Assessment | -79.71$^{***}$ | 31.18$^{***}$ | 21.75$^{***}$ | 25.98$^{***}$ | 19.15$^{***}$ |
| | (2.16) | (4.24) | (3.74) | (3.83) | (5.38) |
| Female (0/1) | | | -8.39$^{***}$ | -8.53$^{***}$ | -8.18$^{***}$ |
| | | | (0.51) | (0.52) | (0.61) |
| ESCS | | | 31.81$^{***}$ | 30.45$^{***}$ | 27.44$^{***}$ |
| | | | (0.32) | (0.32) | (0.36) |
| Large City (0/1) | | | | 2.90$^{***}$ | 7.09$^{***}$ |
| | | | | (0.95) | (1.15) |
| Private School (0/1) | | | | 7.07$^{***}$ | 7.98$^{***}$ |
| | | | | (1.32) | (1.50) |
| Number of Students | | | | 0.01$^{***}$ | 0.01$^{***}$ |
| | | | | (0.00) | (0.00) |
| Whether Academic Achievement is Considered in Admissions (0/1) | | | | 9.60$^{***}$ | 11.14$^{***}$ |
| | | | | (1.11) | (1.25) |
| GDP per Capita based on PPP (log transformed) | | | | | 12.50$^{*}$ |
| | | | | | (6.61) |
| Educational Expenditures as a % of GDP | | | | | 11.81$^{***}$ |
| | | | | | (1.20) |
| Monitoring School's Progress (Accountability) | | | | | -26.45$^{**}$ |
| | | | | | (10.95) |
| Year fixed effects | Y | Y | Y | Y | Y |
| Country fixed effects | | Y | Y | Y | Y |
| Constant | 498.52$^{***}$ | 352.84$^{***}$ | 403.16$^{***}$ | 385.60$^{***}$ | 263.07$^{***}$ |
| | (2.15) | (3.83) | (3.49) | (3.75) | (59.29) |
| Observations | 1,304,482 | 1,304,482 | 1,304,457 | 1,303,445 | 997,114 |
| $R^2$ | 0.03 | 0.32 | 0.42 | 0.42 | 0.43 |

Note: Standard errors in parentheses. Survey weights incorporated. Missing data for independent variables besides the key independent variable and the country covariates are imputed using single imputation.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

Models 2-5 have country and year fixed effects. They show a positive and statistically significant relationship between incorporating student assessment data in teacher evaluations and student math achievement. Going from no schools to 100 percent of schools incorporating student assessment data in their teacher evaluations predicts about a quarter of a standard deviation increase in student test scores in math. The median positive change over time of incorporating student assessment data in teacher evaluations is 16 percentage points. Using the range of estimates from models 2-5, this relates to an effect of 3 to 5 percent (95% CI: 1 to 6 percent) of a standard deviation for math. This might seem small, but it is a substantive effect, particularly for this age group. According to Hill, Bloom, Black, and Lipsey (2008), the average gain for an American 10th or 11th grader (the age for the PISA sample) in a year is 14 percent of a standard deviation in math and 19 percent of a standard deviation in reading. The inclusion of country-level covariates does not substantially affect the key estimate.

For reading in Table 5, the Model 1 results show the same negative pattern, again suggesting that the country-years with lower student achievement are the ones that more often use this type of teacher evaluation system. Models 2-5 are smaller and either nonsignificant or marginally significant, which suggests that the effects of using this type of evaluation system are subject-specific. With the median positive change over time of incorporating student assessment data in teacher evaluations of 16 percentage points, the reading results translate to an effect of -0.5 to 1.2 percent (95% CI: -2 to 3 percent) of a standard deviation change in reading test scores.

Table 5: Average Student Achievement Scores on a Reading Test, Results on Pooled Sample

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Teacher Evaluation using Student Assessment | -86.85$^{***}$ | 7.48$^{*}$ | -2.90 | 2.64 | 7.98 |
| | (2.13) | (4.31) | (3.70) | (3.75) | (5.12) |
| Female (0/1) | | | 33.14$^{***}$ | 32.97$^{***}$ | 32.63$^{***}$ |
| | | | (0.49) | (0.49) | (0.59) |
| ESCS | | | 31.41$^{***}$ | 29.60$^{***}$ | 26.58$^{***}$ |
| | | | (0.29) | (0.30) | (0.29) |
| Large City (0/1) | | | | 6.33$^{***}$ | 10.79$^{***}$ |
| | | | | (0.90) | (1.06) |
| Private School (0/1) | | | | 8.56$^{***}$ | 9.34$^{***}$ |
| | | | | (1.23) | (1.37) |
| Number of Students | | | | 0.01$^{***}$ | 0.01$^{***}$ |
| | | | | (0.00) | (0.00) |
| Whether Academic Achievement is Considered in Admissions (0/1) | | | | 10.20$^{***}$ | 11.96$^{***}$ |
| | | | | (1.05) | (1.14) |
| GDP per Capita based on PPP (log transformed) | | | | | 22.37$^{***}$ |
| | | | | | (5.66) |
| Educational Expenditures as a % of GDP | | | | | 1.81 |
| | | | | | (1.21) |
| Monitoring School's Progress (Accountability) | | | | | -23.71$^{**}$ |
| | | | | | (10.79) |
| Year fixed effects | Y | Y | Y | Y | Y |
| Country fixed effects | | Y | Y | Y | Y |
| Constant | 511.83$^{***}$ | 367.85$^{***}$ | 397.74$^{***}$ | 376.08$^{***}$ | 161.59$^{***}$ |
| | (1.95) | (4.03) | (3.67) | (3.86) | (51.08) |
| Observations | 1,304,482 | 1,304,482 | 1,304,457 | 1,303,445 | 997,114 |
| $R^2$ | 0.04 | 0.23 | 0.35 | 0.36 | 0.36 |

Note: Standard errors in parentheses. Survey weights incorporated.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

*Additional Tests*

There are two sets of interactions that further explain the relationship between using a teacher evaluation system that incorporates student assessment data and student achievement. Table 6 shows the results of these interactions. Models 1 and 2 have an interaction between the average achievement of a country-year and the teacher evaluation system. The estimates for both math and reading are marginally significant or not significant, but the direction suggests that the teacher evaluation system might be more or only effective for countries that have lower average achievement.

Table 6: Interactions with the Teacher Evaluation System

| | Mathematics | | | Reading | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| Teacher Evaluation using Student Assessment | 39.07 | 36.04 | 878.10$^{***}$ | 39.18 | 48.79 | 334.57$^{***}$ |
| | (26.01) | (29.07) | (79.42) | (35.16) | (40.11) | (75.52) |
| Average Score (country) | 1.00$^{***}$ | 1.03$^{***}$ | | 1.04$^{***}$ | 1.10$^{***}$ | |
| | (0.04) | (0.05) | | (0.06) | (0.07) | |
| Interaction of Teacher Eval and Average Score (country) | -0.09$^{*}$ | -0.08 | | -0.09 | -0.10 | |
| | (0.05) | (0.06) | | (0.07) | (0.08) | |
| GDP per capita (logged) | | | 79.06$^{***}$ | | | 47.68$^{***}$ |
| | | | (6.74) | | | (6.96) |
| Interaction of Teacher Eval and GDP per Capita (logged) | | | -86.67$^{***}$ | | | -32.95$^{***}$ |
| | | | (7.73) | | | (7.36) |
| Student & School Covariates | Y | Y | Y | Y | Y | Y |
| Country Covariates | | Y | Y | | Y | Y |
| Fixed Effects | Y | Y | Y | Y | Y | Y |
| Observations | 1,303,445 | 997,114 | 997,114 | 1,303,445 | 997,114 | 997,114 |
| $R^2$ | 0.43 | 0.44 | 0.44 | 0.36 | 0.37 | 0.36 |

Note: Standard errors in parentheses. Survey weights incorporated.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

Model 3 has an interaction between teacher evaluation and logged GDP per capita. The estimates are statistically significant and negative for both math and reading, which suggests that the teacher evaluation model is more effective for lesser developed countries. This is unsurprising compared to the results in Models 1 and 2, given the known positive relationship between GDP per capita and student test scores (e.g., Hanushek et al., 2013). For both math and reading, the relationship between teacher evaluation and student achievement turns from positive to negative around a logged GDP per capita of 10.15, which is around the average logged GDP per capita of the Czech Republic, South Korea, and Portugal. At a higher logged GDP per capita, the relationship between the teacher evaluation system and student achievement is negative, which means that having a teacher evaluation system that includes student assessment data predicts lower student achievement, on average.

The main results and the results from Table 6 require that a country has participated in at least two of the four waves of PISA (2000, 2003, 2009, 2012) because of the country fixed effects. Table 7 shows the results of tests on a more restricted sample that requires a country to have participated in three or four of the waves. Neither the math nor the reading estimates change much when I restrict the sample in this way. The math key estimates are positive, statistically significant, and close to an effect of a quarter of a standard deviation. The reading key estimates are non-significant or positive and marginally significant, as are the main estimates.

Table 7: Robustness Tests Based on Minimum Number of Waves per Country

| | Mathematics | | Reading | |
|---|---|---|---|---|
| Teacher Evaluation using Student Assessment: **Three Waves** | 26.48$^{***}$ | 19.14$^{***}$ | 2.93 | 8.70$^{*}$ |
| | (3.87) | (5.39) | (3.78) | (5.13) |
| Observations | 1,126,938 | 888,747 | 1,126,938 | 888,747 |
| Teacher Evaluation using Student Assessment: **Four Waves** | 29.29$^{***}$ | 19.87$^{***}$ | -0.13 | 7.59 |
| | (4.17) | (5.69) | (4.05) | (5.38) |
| Observations | 959,011 | 755,866 | 959,011 | 755,866 |
| Country Covariates | | Y | | Y |

Note: Non-country covariates, country fixed effects, and year fixed effects are included in each model. Standard errors in parentheses.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

The results in Table 8 replicate the primary analysis in my paper except that it is at the country-year level. I include country and year fixed effects and country covariates (in some models) but not student- or school-level covariates. The results for the math models are no longer statistically significant and are either close to zero or negative. The sample size is drastically smaller than in the main analysis, which partially explains the result. The other major difference is how the countries are weighted in this analysis. In the main analysis, countries are weighted by the size of their student population and the number of waves in which the country participated, so countries with a greater number of students represented in the analysis have more weight. In this country-year level analysis, countries are weighted equally, so the average effect has changed to be zero or perhaps negative. I conclude that the positive effect from my main analysis is more prevalent in country-years with larger student populations, in addition to the finding that it is more prevalent in country-years with a lower GDP per capita. The results from the reading models are similar to what I have shown for other reading models, close to zero or perhaps slightly positive.

Table 8: Country-Year Level Analysis of Teacher Evaluation System

| | Mathematics | | Reading | |
|---|---|---|---|---|
| Teacher Evaluation using Student Assessment | 0.89 | -14.52 | 8.19 | -3.38 |
| | (10.43) | (15.07) | (9.62) | (13.16) |
| Country Covariates | | Y | | Y |
| Observations | 196 | 153 | 196 | 153 |

Note: Country and year fixed effects are included in each model. Standard errors in parentheses.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

Table 9 displays the results of a country-level growth analysis. The PISA test scores are re-normed every few years, so the scores are only directly comparable for certain waves: 2003 to 2009 and 2012 for math and 2000 to 2003 and 2009 to 2012 for reading. I use a model that regresses the change in average test scores on the change in the percent of students exposed to the teacher evaluation system of interest within each country from wave to the next. I do not include any covariates or fixed effects in the model. As with the results in Table 8, the small sample size has decreased the precision of these estimates drastically. All estimates except one are not statistically significant, but most are positive.

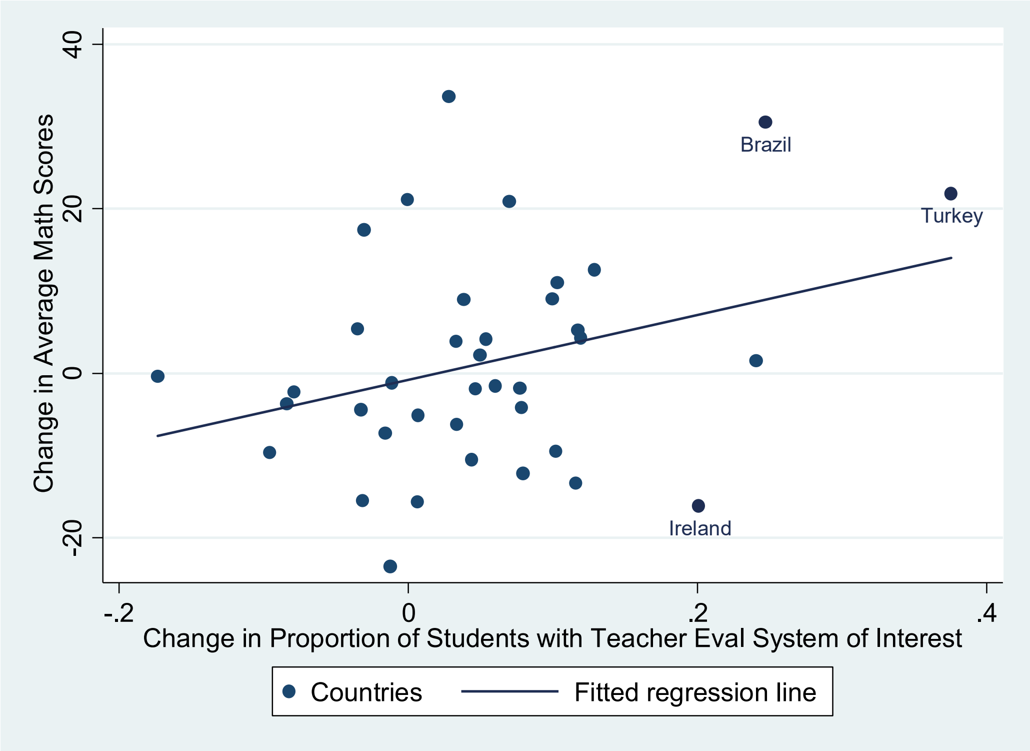
Table 9: Student Achievement and Teacher Evaluation System Growth

| | Mathematics | Reading |
|---|---|---|
| Change in Teacher Evaluation using Student Assessment: **2000 to 2003** | - | 22.83 |
| | - | (16.74) |
| Observations | - | 33 |
| Change in Teacher Evaluation using Student Assessment: **2003 to 2009** | 39.49* | - |
| | (20.28) | - |
| Observations | 38 | - |
| Change in Teacher Evaluation using Student Assessment: **2003 to 2012** | 22.52 | - |
| | (22.03) | - |
| Observations | 38 | - |
| Change in Teacher Evaluation using Student Assessment: **2009 to 2012** | -9.22 | 0.64 |
| | (16.42) | (16.64) |
| Observations | 59 | 59 |

Note: Standard errors in parentheses.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

The marginally significant result from the 2003 to 2009 math growth analysis says that a 100 percent increase in the use of student assessment data in teacher evaluations leads to an increase of 38 percent of a standard deviation. The other, smaller estimates and the fact that the countries are weighted equally in this analysis lead me to conclude that some of the countries that participated in both the 2003 and 2009 waves are unique in some way. Figure 3 is a scatterplot of this growth model, and it shows that some countries have particularly high leverage over the regression line. For the 2003 to 2009 change, Turkey, Ireland, and Brazil have large, positive changes in their use of the teacher evaluation system and a value for Cook's distance (an estimate of the influence of each data point on a least-squares regression analysis) above the traditional cut off of 4/*n*. Turkey and Brazil both have large, positive changes in their student test scores as well, while Ireland has a negative change. Perhaps these countries have education systems that are particularly conducive (or unconducive, for Ireland) to a teacher evaluation system that incorporates student assessment data or their teacher evaluation systems are particularly adept at increasing (or decreasing) student learning. Future research should explore these possibilities in more depth to disentangle characteristics of teacher evaluation systems from characteristics of the broader country.

Figure 3: Mathematics Student Achievement and Teacher Evaluation System Growth, 2003 to 2009



## Discussion & Conclusion

I designed this study to test the idea that comes from new institutionalism that incorporating student assessment data into teacher evaluations will increase student learning, measured using the low-stakes PISA assessment. I find that the notion is supported on average among included OECD and partner countries for the PISA age group in mathematics but not reading. This effect is more prevalent in country-years with a low GDP per capita, and the positive relationship is reversed, on average, for country-years with a higher GDP per capita for both math and reading. Country-years with a higher GDP per capita experience negative student learning, on average, if they more frequently include student test scores in teacher evaluations. Looking at this result through an education production function lens, I conclude that this type of evaluation system is actually harmful for student learning; however, it can make up for inputs that positively affect learning but are lacking in lesser developed countries. Once those other positive inputs are in place, as a country develops, this type of teacher evaluation system no longer provides benefits and instead negatively affects the education system. The key relationship is non-significant in most models for reading achievement, though the interaction with GDP per capita persists at a lower rate. Student population size is possibly also important, as my main results weight countries by student population size, though I do not look in-depth at this relationship. A country-year analysis shows the average effect shrinks or becomes negative when countries are equally weighted. These effects are averages across the sample and are generalizable only to the included OECD and partner countries. Importantly, my sample excludes many developing countries, primarily from Africa and South America.

Due to our global testing culture, high-stakes standardized testing is frequently assumed and accepted as key to the development of education systems (Smith, 2016), and the results of my study show that these assessments can positively affect student learning, but there are concerns with the use of this type of system. Not all education systems have the financial or technical capacity to develop appropriate tests, test students frequently enough for their scores to be used in an evaluation system, store the test data, and analyze the data to reliably and validly estimate the impact of teachers on their students (Best et al., 2013; Isoré, 2009; Kamens & McNeely, 2010). The test itself can be poorly written or not designed with the ultimate purpose to evaluate teachers, and all tests include measurement error (Hanushek & Rivkin, 2010; Isoré, 2009; Kane, Staiger, Grissmer, & Ladd, 2002). Attaching high-stakes decisions to the tests, like compensation or hiring and firing decisions, provides incentives for cheating, teaching narrowly to the test, and foregoing the teaching of non-tested subjects (Hanushek & Rivkin, 2010; Isoré, 2009). There are additional concerns about using value-added systems, including how teacher value-added is calculated and whether teacher effects persist over time (Hanushek & Rivkin, 2010; Isoré, 2009; Kane et al., 2002).

The use of student test scores to evaluate teachers can increase stress on teachers who are subject to this type of evaluation system, and this stress can lead to teacher attrition, which is already a significant issue in some countries (OECD, 2005). Liu and Onwuegbuzie (2012) share data collected from 510 teachers in the Jilin Province of China. The surveyed teachers report that student test scores are very important for teacher evaluations, particularly for teachers whose students are at the end of middle school. The high expectations of principals (and parents) for students to get high test scores put teachers under substantial pressure, which causes a lot of stress, sometimes manifested in physical symptoms, while teachers prepare their students for exams and wait for the results. Perryman and Calvert (2019) survey early career teachers in London and find that many teachers either leave or considered leaving teaching because of accountability pressures and related academic targets. Additionally, Smith and Kubacka (2017) find, using data from the 2013 wave of TALIS, that teachers feel that feedback has a more limited impact on their instruction when it emphasizes student achievement.

While I was unable to include teacher working conditions, attrition, or the usefulness of the feedback in this analysis due to data limitations, a clear extension of this research involves further analyzing the relationship between this type of evaluation system and teacher outcomes on a similar international scale.

The theory of new institutionalism can explain why the incorporation of student assessment data in teacher evaluations has persisted in many OECD and partner countries throughout the sample years, despite the growing evidence of its negative effects. According to new institutionalism, the goal of institutions is not just to maximize efficiency and profits (in this case, student test scores) but to survive and gain legitimacy. To do so, institutions like education systems often must bend to isomorphic pressures of peer institutions in ways such as copying teacher evaluation systems from other countries. These pressures are influential, despite possible evidence that they result in a system that does not work well (Meyer & Rowan, 2006). All regions except Europe increased their use of the teacher evaluation system examined in my paper over time, with possible leveling off as of 2009. The international pressure to use this type of evaluation system for education systems to retain their legitimacy explains why many education systems continue to use and even expand this type of evaluation system, despite evidence of its negative effects.

Another extension to this paper would be to take a closer look at what this type of evaluation system means in practice. I cannot know from the questionnaire the extent to which the student data are incorporated into the teachers' evaluations (shown to teacher for self-evaluation, assigned a specific percent to create an evaluation score, etc.). It is not possible from my analysis to identify what specific teacher evaluation methods are most productive in improving student achievement. Further research could shed light on this question and provide insight on how to include test scores in teacher evaluations while minimizing concerns about teacher working conditions and attrition.

Cross-country research can help identify patterns that are masked in within-country systems for reasons such as the simultaneous introduction of related policies with the policy of interest. However, it is challenging to identify causal effects using international, cross-sectional data. I use an identification strategy that removes the most egregious mechanisms for bias and attempts to control for other time-varying country factors that might bias the estimates; my results are consistent across robustness tests that affect the sample of countries. It is still possible that I excluded other important aspects that might bias the results. These concerns, among others, make it particularly important to consider these results in conjunction with other research to provide additional detail on contextual effects and other components of teacher evaluation systems. For instance, my findings that countries with a high GDP per capita experience a negative relationship, on average, between incorporating student test scores in teacher evaluations and student learning do not explain why Adnot, Dee, Katz, and Wyckoff (2017) found improved student performance in a city in the high-income United States. Another avenue to consider is student population size, as my analysis weights countries by their student population size, and instead weight countries equally.

An obvious and primary goal of education systems is to raise student achievement, but it is not the only goal. The potential negative effects of incorporating student assessment data into teacher evaluations are clear, but so are the potential positive ones. There are ways to maximize student learning that also allow students to thrive within and outside traditionally tested academic subjects. There are ways to measure teacher performance accurately without incentivizing harmful effects, like cheating or negative labor market outcomes. My study is one piece of the broader puzzle of how to best build an effective teacher evaluation system.

# REFERENCES

Addey, C., & Sellar, S. (2018). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, M. Novelli, & H. K. Altinyelken (Eds.), *Global Education Policy and International Development: New Agendas, Issues and Policies*.

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. https://doi.org/10.3102/0162373716663646

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher Quality, Opportunity Gap, and National Achievement in 46 Countries. *Educational Researcher*, *36*(7), 369–387. https://doi.org/10.3102/0013189X07308739

Anderson, J. A. (2005). *Accountability in Education*. Paris. https://doi.org/10.1177/1477878518805308

Benavot, A., & Tanner, E. (2007). *The Growth of National Learning Assessments in the World , 1995-2006*. Retrieved from http://unesdoc.unesco.org/images/0015/001555/155507e.pdf

Best, M., Knight, P., Lietz, P., Lockwood, C., Nugroho, D., & Tobin, M. (2013). *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries*. London.

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work through accountability reforms*. The World Bank. https://doi.org/10.1596/978-0-8213-8679-8

Choi, H. J., & Park, J.-H. (2016). An Analysis of Critical Issues in Korean Teacher Evaluation Systems. *Center for Educational Policy Studies Journal*, *6*(2), 151–171. Retrieved from http://search.proquest.com.ezproxy.puc.cl/education/docview/1811918556/670B1F5383174F5DPQ/1?accountid=16788

Dahler-Larsen, P. (2012). Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable. *Scandinavian Journal of Public Administration*, *16*(3), 29–46.

Danielson, C. (2001). New Trends in Teacher Evaluation. *Educational Leadership*, *58*(5), 12–15.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives*, *8*(1). https://doi.org/10.14507/epaa.v8n1.2000

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, *106*(6), 1047–1085. https://doi.org/10.1111/j.1467-9620.2004.00372.x

Hanberger, A. (2013). Framework for exploring the interplay of governance and evaluation. *Scandinavian Journal of Public Administration*, *16*(3), 9–27.

Hanushek, E. A. (1986). The Economics of Schooling: Prodution and Efficiency in Public Schools. *Journal of Economic Literature*, *24*(3), 1141–1177.

Hanushek, E. A., Link, S., & Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics*, *104*, 212–232. https://doi.org/10.1016/j.jdeveco.2012.08.002

Hanushek, E. A., & Rivkin, S. G. (2010). *Using Value-Added Measures of Teacher Quality*. https://doi.org/10.1257/aer.100.2.267

Hardy, I., & Boyle, C. (2011). My school? Critiquing the abstraction and quantification of education. *Asia-Pacific Journal of Teacher Education*, *39*(3), 211–222. https://doi.org/10.1080/1359866X.2011.588312

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research, *2*(3), 172–177.

IMPACT. (2018). *Teachers (Grades 4+) with Individual Value-Added Student Achievement and Student Survey Data*.

Isoré, M. (2009). *Teacher Evaluation: Current Practices in OECD Countries and a Literature Review* (No. 23). Paris. https://doi.org/10.1787/223283631428

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the Growth of International Educational Testing and National Assessment. *Comparative Education Review*, *54*(1), 5–25. https://doi.org/10.1086/648471

Kane, T. J., Staiger, D., Grissmer, D. O., & Ladd, H. F. (2002). *Volatility in School Test Scores: Implications for Test-Based Accountability Systems*. *Brookings Papers on Education Policy* (Vol. 5). https://doi.org/10.1353/pep.2002.0010

LCSDSN. (2015). *Indicators and a Monitoring Framework for Sustainable Development Goals (SDGs)*.

Leithwood, K., & Earl, L. (2000). Educational Accountability Effects: An International Perspective. *Peabody Journal of Education*, *75*(4), 1–18. https://doi.org/10.1287/isre.l060.0080

Lindgren, L., Hanberger, A., & Lundström, U. (2016). Evaluation systems in a crowded policy space: Implications for local school governance. *Education Inquiry*, *7*(3). https://doi.org/10.3402/edui.v7.30202

Liu, S., & Onwuegbuzie, A. J. (2012). Chinese teachers' work stress and their turnover intention. *International Journal of Educational Research*, (53).

Macdonald, K. (2008). PV: Stata module to perform estimation with plausible values. Statistical Software Components S456951: Boston College Department of Economics.

Martins, P. S. (2009). *Individual teacher incentives, student achievement and grade inflation.*

Meyer, H.-D., & Benavot, A. (2013). PISA and the globalization of education governance: some puzzles and problems. In *PISA, power and policy. The emergence of a global educational governance* (pp. 9–26). Oxford: Symposium Books. https://doi.org/10.1080/0969594X.2013.877874

Meyer, H.-D., & Rowan, B. (2006). Institutional Analysis and the Study of Education. In *The New Institutionalism in Education* (pp. 1–13). Albany: State University of New York Press.

Mortimore, P. (2009). *Alternative models for analysing and representing countries' performance in PISA*. Brussels.

Norlund, A., Marzano, A., & De Angelis, M. (2016). Decentralization tendencies and teacher evaluation policies in European countries. *Italian Journal of Educational Research*, *9*(17).

OECD. (2005). *Teachers Matter: Attracting, developing and retaining effective teachers*. https://doi.org/10.1787/9789264022157-ja

OECD. (2018). *Education Policy Outlook: Mexico*.

Perryman, J., & Calvert, G. (2019). What Motivates People To Teach, and Why Do They Leave? Accountability, Performativity and Teacher Retention. *British Journal of Educational Studies*, 1–21. https://doi.org/10.1080/00071005.2019.1589417

Peterson, K. D. (2000). Problems of Teacher Evaluation. In *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousaand Oaks: Corwin Press.

Pons, A., Amoroso, J., Herczynski, J., Kheyfets, I., Lockheed, M., & Santiago, P. (2015). *OECD Reviews of School Resources: Kazakhstan*. https://doi.org/http://dx.doi.org/10.1787/9789264245891-en

Postlewaite, T. N., & Kellaghan, T. (2008). *National Assessments of Educational Achievement*. https://doi.org/10.1596/978-0-8213-7929-5

Santiago, P., Donaldson, G., Looney, A., & Nusche, D. (2012). *OECD Reviews of Evaluation and Assessment in Education: Portugal*.

Santiago, P., Gilmore, A., Nusche, D., Ravela, P., & Sammons, P. (2012). *OECD Reviews of Evaluation and Assessment in Education: Mexico*.

Shewbridge, C., Ehren, M., Santiago, P., & Tamassia, C. (2012). *OECD Reviews of Evaluation and Assessment in Education: Luxembourg*.

Smith, W. C. (2016). An Introduction to the Global Testing Culture. In *The Global Testing Culture: Shaping Education Policy, Perceptions, and Practice*. Oxford: Symposium Books.

Smith, W. C., & Kubacka, K. (2017). The Emphasis of Student Test Scores in Teacher Appraisal Systems. *Education Policy Analysis Archives*, *25*(86).

Specht, W., & Sobanski, F. (2012). *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes: Country Background Report for Austria*.

Sung Tae Jang. (2016). The Effectiveness of Tying Teacher Evaluation Policy to Student Achievement in South Korea. *US-China Education Review A*, *6*(1), 1–19. https://doi.org/10.17265/2161-623x/2016.01.001

Tatto, M. T. (2006). Education reform and the global regulation of teachers' education, development and work: A cross-cultural analysis. *International Journal of Educational Research*, *45*, 231–241. https://doi.org/10.1016/j.ijer.2007.02.003

Wößmann, L. (2007). International evidence on school competition, autonomy, and accountability: A review. *Peabody Journal of Education*, *82*(2–3), 473–497. https://doi.org/10.1080/01619560701313176

## APPENDIX A: ADDITIONAL 2006 ANALYSIS

In 2006, the principals were asked, "In your school, are achievement data used in any of the following ways?" with one of the situations as, "Achievement data are used in evaluation of teachers' performance." Principals were asked to answer yes or no. Both this question and the one used in 2000, 2003, 2009, and 2012 were asked in 2009, so I used that overlap to compare responses to the two different questions. In 76 percent of the schools that answered both questions, the principal answered the same way. The proportion who answered affirmatively to each question was approximately the same. In most countries with differences in answers of over 10 percentage points, there is a higher affirmation rate for the second version of the question, the one that is in 2006. Because of this, I assumed the country fixed effects would capture any problems that might arise with the different question. Table A1 shows the updated percent of students who attend a school with this evaluation system, including 2006 data.

Table A1: Percent of Students Who Attend a School that Uses Student Assessment Data When Evaluating Teachers

| Country Name | 2000 | 2003 | 2006* | 2009 | 2012 |
|---|---|---|---|---|---|
| Albania | 91% | | | 90% | 87% |
| Argentina | 64% | | 54% | 53% | 51% |
| Australia | 33% | 34% | 43% | 44% | 50% |
| Austria | 19% | 36% | 26% | 26% | 39% |
| Belgium | 26% | 19% | 15% | 31% | 35% |
| Brazil | 46% | 56% | 78% | 80% | 80% |
| Bulgaria | 89% | | 59% | 92% | 93% |
| Canada | 32% | 31% | 19% | 35% | 30% |
| Chile | 55% | | 56% | 58% | 61% |
| Chinese Taipei | | | 30% | 58% | 48% |
| Colombia | | | 73% | 63% | 60% |
| Costa Rica | | | | 57% | 71% |
| Croatia | | | 39% | 55% | 56% |
| Czech Republic | 67% | 62% | 91% | 60% | 63% |
| Denmark | 4% | 4% | 22% | 8% | 27% |
| Estonia | | | 86% | 72% | 65% |
| Finland | 38% | 32% | 14% | 24% | 16% |
| Germany | 12% | 12% | 28% | 22% | 24% |
| Greece | 18% | 15% | 9% | 22% | 14% |
| Hong Kong | 57% | 64% | 63% | 76% | 80% |
| Hungary | 69% | 77% | 92% | 60% | 58% |
| Iceland | 43% | 31% | 25% | 29% | 39% |
| Indonesia | 97% | 87% | 97% | 98% | 96% |
| Ireland | 29% | 17% | 30% | 37% | 47% |

| | | | | | |
|---|---|---|---|---|---|
| Israel | 82% | | 94% | 85% | 82% |
| Italy | 86% | 23% | 25% | 20% | 30% |
| Japan | 81% | 82% | 26% | 78% | 76% |
| Jordan | | | 82% | 82% | 72% |
| Kazakhstan | | | | 99% | 100% |
| Latvia | 86% | 86% | 91% | 92% | 93% |
| Lithuania | | | 84% | 71% | 74% |
| Luxembourg | 0% | 21% | 5% | 22% | 22% |
| Macao (China) | | 82% | 41% | 74% | 75% |
| Malaysia | | | | 92% | 92% |
| Mexico | 77% | 77% | 83% | 80% | 77% |
| Montenegro | | | 71% | 56% | 92% |
| Netherlands | 47% | 42% | 73% | 50% | 68% |
| New Zealand | 48% | 53% | 47% | 61% | 68% |
| Norway | 10% | 19% | 40% | 24% | 30% |
| Peru | 78% | | | 80% | 78% |
| Poland | 93% | 73% | 89% | 79% | 79% |
| Portugal | 24% | 35% | 39% | 35% | 50% |
| Qatar | | | 93% | 84% | 87% |
| Romania | 20% | | 97% | 85% | 75% |
| Russia | 99% | 99% | 100% | 98% | 99% |
| Singapore | | | | 85% | 88% |
| Slovak Republic | | 75% | 75% | 80% | 69% |
| Slovenia | | | 27% | 40% | 38% |
| South Korea | 14% | 54% | 34% | 66% | 85% |
| Spain | 42% | 36% | 42% | 44% | 50% |
| Sweden | 12% | 21% | 49% | 22% | 44% |
| Switzerland | 6% | 37% | 8% | 41% | 36% |
| Thailand | 76% | 71% | 86% | 95% | 91% |
| Tunisia | | 63% | 82% | 76% | 67% |
| Turkey | | 34% | 75% | 71% | 71% |
| United Arab Emirates | | | | 91% | 94% |
| United Kingdom | 78% | 86% | 94% | 83% | 88% |
| United States | 56% | 55% | 42% | 58% | 60% |
| Uruguay | | 41% | 46% | 37% | 31% |

Note: Survey weights incorporated.

* 2006 percentages are sometimes higher than the trend would suggest due to the different survey question.

The 2006, 2009, and 2012 waves have questions similar to the accountability/monitoring question as well: "In your school, are achievement data used in any of the following ways? Achievement data are tracked over time by an administrative authority" with a yes/no response option. I used the overlap of questions in year 2009 and 2012 to compare the answers. For both years, 74-75 percent of the answers were consistent across the two questions, and, in almost all country-years, the mean response to the monitor question was higher than in the track question, meaning the year fixed effects should help adjust for this pattern. I use the monitor question in 2000, 2003, 2009, and 2012 and the track question in 2006. Table A2 shows descriptive statistics for the pooled sample with 2006 included.

Table A2: Descriptive Statistics on Pooled Sample with 2006 Included

| | Mean/Percent | Sample Standard Deviation | Sample N |
|---|---|---|---|
| *The average student has…* | | | |
| **Math Score** | **452** | **104.94** | **1,682,291** |
| **Reading Score** | **457** | **104.05** | **1,676,680** |
| ESCS (continuous, ranged -7.5 to 5.1) | -.74 | 1.19 | 1,653,209 |
| *X% of students are…* | | | |
| Female | 50.4% | - | 1,680,587 |
| *X% of students go to a school that is…* | | | |
| In a Large City | 40.2% | - | 1,621,255 |
| Private | 18.0% | - | 1,600,159 |
| Has Admission Based on Prior Academics | 56.3% | - | 1,626,740 |
| Monitoring School's Progress (*Accountability*) | 82.9% | - | 1,682,291 |
| **Evaluating Teachers using Student Data** | **67.0%** | - | **1,624,535** |
| *The average student goes to a school that has…* | | | |
| Number of Students | 806.6 | 639.72 | 1,580,091 |
| *The average student is in a country with…* | | | |
| GDP per Capita Based on Purchasing Power Parity (constant 2011 international $) | $22,247 | $19,318.62 | 1,661,599 |
| GDP per Capita (log transformed) | 9.99 | 10.21 | 1,661,599 |
| Education Expenditures as a % of GDP | 4.4% | 1.10% | 1,277,057 |

Note: Survey weights incorporated in the mean/percent. Bolded values are the dependent variables and the key variable of interest.

Tables A3 and A4 show the results for the analyses including 2006. While the overall conclusions do not change substantively, the point estimates do differ enough for the key variable of interest and the *Accountability* variable from the analyses including 2006 to those excluding 2006 that I decided to exclude 2006 from the primary analysis.

Table A3: Mathematics Results on Pooled Sample with 2006

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Teacher Evaluation using Student Assessment | -88.20$^{***}$ | 29.04$^{***}$ | 23.14$^{***}$ | 25.36$^{***}$ | 21.60$^{***}$ |
| | (1.76) | (3.62) | (3.14) | (3.17) | (4.32) |
| Female (0/1) | | | -8.94$^{***}$ | -9.10$^{***}$ | -8.77$^{***}$ |
| | | | (0.47) | (0.48) | (0.51) |
| ESCS | | | 31.83$^{***}$ | 30.38$^{***}$ | 27.50$^{***}$ |
| | | | (0.27) | (0.29) | (0.33) |
| Large City (0/1) | | | | 2.79$^{***}$ | 6.85$^{***}$ |
| | | | | (1.05) | (0.95) |
| Private School (0/1) | | | | 6.66$^{***}$ | 9.32$^{***}$ |
| | | | | (1.16) | (1.32) |
| Number of Students | | | | 0.01$^{***}$ | 0.01$^{***}$ |
| | | | | (0.00) | (0.00) |
| Whether Academic Achievement is Considered in Admissions (0/1) | | | | 10.58$^{***}$ | 11.53$^{***}$ |
| | | | | (0.98) | (1.11) |
| GDP per Capita based on PPP (log transformed) | | | | | 6.65 |
| | | | | | (6.57) |
| Educational Expenditures as a % of GDP | | | | | 9.85$^{***}$ |
| | | | | | (1.04) |
| Monitoring School's Progress (Accountability) | | | | | -8.10 |
| | | | | | (6.08) |
| Year fixed effects | Y | Y | Y | Y | Y |
| Country fixed effects | | Y | Y | Y | Y |
| Constant | 503.89$^{***}$ | 354.53$^{***}$ | 402.31$^{***}$ | 385.01$^{***}$ | 299.07$^{***}$ |
| | (2.01) | (3.39) | (2.96) | (3.09) | (57.44) |
| Observations | 1,682,291 | 1,682,291 | 1,682,259 | 1,680,633 | 1,275,922 |
| $R^2$ | 0.04 | 0.31 | 0.41 | 0.41 | 0.43 |

Note: Standard errors in parentheses. Survey weights incorporated.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

Table A4: Reading Results on Pooled Sample with 2006

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Teacher Evaluation using Student Assessment | -91.73$^{***}$ | 7.42$^{*}$ | 2.01 | 5.45 | 9.73$^{**}$ |
| | (1.71) | (3.82) | (4.01) | (3.90) | (4.45) |
| Female (0/1) | | | 33.62$^{***}$ | 33.44$^{***}$ | 33.65$^{***}$ |
| | | | (0.47) | (0.47) | (0.51) |
| ESCS | | | 31.41$^{***}$ | 29.40$^{***}$ | 26.87$^{***}$ |
| | | | (0.29) | (0.34) | (0.29) |
| Large City (0/1) | | | | 6.69$^{***}$ | 9.95$^{***}$ |
| | | | | (1.11) | (0.93) |
| Private School (0/1) | | | | 8.08$^{***}$ | 10.58$^{***}$ |
| | | | | (1.13) | (1.26) |
| Number of Students | | | | 0.01$^{***}$ | 0.01$^{***}$ |
| | | | | (0.00) | (0.00) |
| Whether Academic Achievement is Considered in Admissions (0/1) | | | | 11.03$^{***}$ | 12.31$^{***}$ |
| | | | | (0.93) | (1.03) |
| GDP per Capita based on PPP (log transformed) | | | | | 13.90$^{**}$ |
| | | | | | (5.75) |
| Educational Expenditures as a % of GDP | | | | | 0.80 |
| | | | | | (1.11) |
| Monitoring School's Progress (Accountability) | | | | | -7.57 |
| | | | | | (6.09) |
| Year fixed effects | Y | Y | Y | Y | Y |
| Country fixed effects | | Y | Y | Y | Y |
| Constant | 514.91$^{***}$ | 367.97$^{***}$ | 393.58$^{***}$ | 371.67$^{***}$ | 219.41$^{***}$ |
| | (1.75) | (3.87) | (3.88) | (3.75) | (50.16) |
| Observations | 1,676,680 | 1,676,680 | 1,676,648 | 1,675,022 | 1,275,922 |
| $R^2$ | 0.05 | 0.22 | 0.34 | 0.35 | 0.35 |

Note: Standard errors in parentheses. Survey weights incorporated.
$^{*}$ $p < 0.1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

## APPENDIX B: DESCRIPTIVE STATISTICS

Table B1: Descriptive Statistics by Country (Mean/Percent) – Student Level

| | Math Score | Reading Score | % Female | ESCS |
|---|---|---|---|---|
| Albania | 385.74 | 380.19 | 49.37% | -0.97 |
| Argentina | 387.56 | 403.10 | 53.68% | -0.89 |
| Australia | 518.56 | 519.53 | 48.66% | 0.04 |
| Austria | 505.15 | 488.89 | 50.74% | -0.21 |
| Belgium | 519.11 | 507.12 | 48.62% | -0.05 |
| Brazil | 363.44 | 401.54 | 52.97% | -1.34 |
| Bulgaria | 431.42 | 430.63 | 48.29% | -0.46 |
| Canada | 527.63 | 527.49 | 50.14% | 0.33 |
| Chile | 409.58 | 433.82 | 51.07% | -0.88 |
| Chinese Taipei | 551.67 | 508.88 | 50.21% | -0.41 |
| Colombia | 378.47 | 408.00 | 52.65% | -1.26 |
| Costa Rica | 408.45 | 442.30 | 53.04% | -1.04 |
| Croatia | 465.44 | 480.42 | 48.04% | -0.32 |
| Czech Republic | 501.71 | 488.45 | 49.16% | -0.32 |
| Denmark | 507.15 | 494.91 | 50.02% | 0.29 |
| Estonia | 516.47 | 508.71 | 49.30% | -0.09 |
| Finland | 535.00 | 537.98 | 50.13% | 0.10 |
| Germany | 504.56 | 494.03 | 49.64% | 0.00 |
| Greece | 451.70 | 476.75 | 50.77% | -0.43 |
| Hong Kong | 556.83 | 529.78 | 48.22% | -1.02 |
| Hungary | 487.05 | 486.55 | 49.38% | -0.51 |
| Iceland | 506.72 | 494.85 | 49.52% | 0.45 |
| Indonesia | 369.14 | 388.40 | 50.12% | -1.92 |
| Ireland | 498.84 | 515.81 | 49.82% | -0.24 |
| Israel | 450.23 | 471.94 | 52.90% | -0.04 |
| Italy | 473.23 | 484.77 | 49.73% | -0.40 |
| Japan | 539.99 | 519.90 | 49.60% | -0.37 |
| Kazakhstan | 417.16 | 391.76 | 49.69% | -0.45 |
| Jordan | 385.95 | 401.98 | 50.13% | -0.52 |
| Korea | 547.40 | 533.11 | 44.78% | -0.41 |
| Latvia | 478.14 | 477.02 | 50.92% | -0.54 |
| Lithuania | 477.54 | 472.37 | 49.44% | -0.26 |
| Luxembourg | 480.80 | 470.82 | 49.53% | -0.12 |
| Macao (China) | 529.96 | 497.66 | 49.93% | -1.13 |
| Malaysia | 412.24 | 406.07 | 51.26% | -0.67 |
| Mexico | 403.12 | 417.96 | 50.89% | -1.44 |
| Montenegro | 405.69 | 415.10 | 49.40% | -0.39 |
| Netherlands | 536.82 | 516.54 | 49.38% | 0.01 |
| New Zealand | 518.86 | 519.86 | 49.21% | -0.05 |
| Norway | 494.94 | 502.15 | 49.32% | 0.33 |
| Peru | 348.76 | 363.87 | 50.39% | -1.40 |
| Poland | 491.29 | 496.43 | 50.10% | -0.62 |
| Portugal | 472.93 | 480.08 | 51.34% | -0.84 |

| | | | | |
|---|---|---|---|---|
| Qatar | 372.53 | 380.04 | 48.71% | 0.42 |
| Romania | 431.71 | 430.26 | 51.79% | -0.94 |
| Russian Federation | 473.27 | 456.12 | 50.00% | -0.50 |
| Singapore | 567.55 | 533.86 | 49.06% | -0.23 |
| Slovak Republic | 492.92 | 470.19 | 49.05% | -0.38 |
| Slovenia | 501.73 | 482.11 | 48.62% | -0.05 |
| Spain | 481.85 | 485.68 | 50.09% | -0.24 |
| Sweden | 497.83 | 502.64 | 49.64% | 0.16 |
| Switzerland | 529.99 | 501.36 | 49.25% | -0.06 |
| Thailand | 422.99 | 427.90 | 56.35% | -1.73 |
| Tunisia | 371.69 | 392.56 | 52.08% | -1.61 |
| Turkey | 442.05 | 463.47 | 48.05% | -1.61 |
| United Arab Emirates | 427.54 | 436.71 | 50.58% | 0.25 |
| United Kingdom | 505.58 | 505.68 | 51.36% | -0.06 |
| United States | 485.89 | 499.04 | 49.48% | 0.09 |
| Uruguay | 418.72 | 423.07 | 52.48% | -0.94 |

Note: Survey weights incorporated.

Table B2: Descriptive Statistics by Country (Mean/Percent) – School Level

| | % Students whose school is in a large city | % Students in a private school | % Students in a school that has Admission Based on Academics | Avg. Student goes to a school with X% Funding from Gov | Avg. Student goes to a school with X Teachers | Avg. Student goes to a school with X Students |
|---|---|---|---|---|---|---|
| Albania | 27.69% | 8.18% | 68.58% | 74.05% | 31.24 | 316.15 |
| Argentina | 39.74% | 36.00% | 35.10% | 62.03% | 51.40 | 416.16 |
| Australia | 63.35% | 40.79% | 61.43% | 71.19% | 64.14 | 724.12 |
| Austria | 31.75% | 10.43% | 76.99% | 90.22% | 49.53 | 444.50 |
| Belgium | 21.87% | 70.05% | 59.32% | 87.64% | 62.19 | 547.57 |
| Brazil | 51.67% | 14.44% | 25.54% | 81.84% | 29.46 | 1,076.25 |
| Bulgaria | 42.08% | 1.13% | 92.91% | 95.96% | 48.86 | 494.81 |
| Canada | 49.87% | 7.48% | 54.82% | 89.77% | 61.51 | 748.33 |
| Chile | 58.91% | 55.99% | 68.40% | 72.30% | 35.60 | 704.90 |
| Chinese Taipei | 61.36% | 35.73% | 77.97% | 64.00% | 128.01 | 1,686.06 |
| Colombia | 51.63% | 17.42% | 68.35% | 76.13% | 51.64 | 812.89 |
| Costa Rica | 16.27% | 15.23% | 57.27% | 75.06% | 40.15 | 478.75 |
| Croatia | 36.72% | 1.81% | 99.46% | 94.09% | 44.54 | 401.97 |
| Czech Republic | 24.56% | 6.00% | 69.43% | 94.87% | 30.24 | 378.02 |
| Denmark | 15.52% | 23.46% | 21.81% | 92.45% | 36.31 | 331.16 |
| Estonia | 28.14% | 3.56% | 72.28% | 97.68% | 39.31 | 294.01 |
| Finland | 25.57% | 4.13% | 16.59% | 99.81% | 33.83 | 320.40 |
| Germany | 26.97% | 6.02% | 66.27% | 96.50% | 31.75 | 524.86 |
| Greece | 28.53% | 5.79% | 21.39% | 84.86% | 29.49 | 263.35 |
| Hong Kong | 94.85% | 70.94% | 98.98% | 90.25% | 60.08 | 816.35 |
| Hungary | 41.71% | 11.39% | 93.00% | 90.92% | 46.84 | 412.10 |
| Iceland | 27.44% | 0.68% | 11.76% | 99.51% | 32.60 | 319.35 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Indonesia | 22.31% | 44.09% | 80.43% | 49.21% | 28.97 | 483.01 |
| Ireland | 27.36% | 60.46% | 25.08% | 90.62% | 37.09 | 477.53 |
| Israel | 36.08% | 12.23% | 75.46% | 77.13% | 54.97 | 527.14 |
| Italy | 30.72% | 5.43% | 57.73% | 67.59% | 74.97 | 597.59 |
| Japan | 66.19% | 29.14% | 99.35% | 73.39% | 57.42 | 728.38 |
| Kazakhstan | 41.10% | 3.02% | 62.71% | 94.79% | 69.02 | 392.63 |
| Jordan | 41.42% | 17.59% | 68.17% | 66.52% | 42.86 | 424.03 |
| Korea | 84.33% | 47.80% | 70.47% | 50.28% | 64.27 | 982.28 |
| Latvia | 30.68% | 1.07% | 65.10% | 96.04% | 43.83 | 464.74 |
| Lithuania | 36.55% | 1.19% | 48.05% | 98.02% | 49.06 | 304.55 |
| Luxembourg | 11.32% | 14.15% | 94.67% | 97.20% | 128.83 | 1,092.06 |
| Macao (China) | 99.88% | 95.67% | 96.08% | 71.27% | 97.66 | 1,350.41 |
| Malaysia | 30.08% | 3.04% | 65.03% | 82.78% | 92.93 | 779.67 |
| Mexico | 42.41% | 13.31% | 66.92% | 41.98% | 20.82 | 689.43 |
| Montenegro | 30.05% | 0.44% | 68.12% | 91.70% | 52.83 | 565.19 |
| Netherlands | 29.68% | 70.77% | 97.45% | 95.76% | 41.06 | 768.82 |
| New Zealand | 51.00% | 5.28% | 46.13% | 78.91% | 65.49 | 861.41 |
| Norway | 16.82% | 1.36% | 7.11% | 99.62% | 26.18 | 238.50 |
| Peru | 39.34% | 20.89% | 48.39% | 43.84% | 39.16 | 560.15 |
| Poland | 27.10% | 2.13% | 63.80% | 95.25% | 33.49 | 384.19 |
| Portugal | 21.20% | 9.46% | 30.30% | 84.39% | 101.88 | 778.51 |
| Qatar | 45.17% | 34.89% | 69.09% | 65.50% | 81.82 | 671.19 |
| Romania | 39.92% | 0.70% | 83.67% | 93.93% | 55.34 | 709.94 |
| Russian Federation | 43.07% | 0.22% | 47.68% | 94.08% | 43.93 | 576.49 |
| Singapore | 100.00% | 2.11% | 98.50% | 81.65% | 87.72 | 820.59 |
| Slovak Republic | 16.38% | 10.30% | 70.89% | 94.79% | 32.09 | 346.21 |
| Slovenia | 38.05% | 2.55% | 68.70% | 95.05% | 31.16 | 290.14 |
| Spain | 40.57% | 35.94% | 11.07% | 85.97% | 55.23 | 556.46 |
| Sweden | 22.65% | 7.94% | 5.97% | 99.85% | 33.79 | 377.89 |
| Switzerland | 15.17% | 6.17% | 74.00% | 94.54% | 28.17 | 438.54 |
| Thailand | 27.96% | 16.18% | 86.96% | 76.63% | 76.29 | 1,323.90 |
| Tunisia | 19.13% | 1.36% | 67.35% | 77.29% | 64.66 | 717.38 |
| Turkey | 56.68% | 1.59% | 64.31% | 61.31% | 46.64 | 665.29 |
| United Arab Emirates | 57.79% | 52.23% | 90.75% | 46.21% | 95.77 | 704.76 |
| United Kingdom | 32.96% | 16.35% | 23.25% | 91.99% | 63.80 | 886.36 |
| United States | 34.57% | 7.16% | 45.59% | 90.53% | 82.61 | 1,055.96 |
| Uruguay | 40.53% | 16.27% | 27.25% | 77.52% | 10.62 | 565.74 |

Note: Survey weights incorporated.

Table B3: Descriptive Statistics by Country (Mean/Percent) – Country-Year Level

| | % Students who Attend School where Teachers Evaluated Using Student Data | Accountability | GDP per Capita Based on PPP | Educational Expenditures as a % of GDP |
|---|---|---|---|---|
| Albania | 89.08% | 95.34% | $8,362.75 | 3.29 |
| Argentina | 55.55% | 81.71% | $16,049.52 | 5.14 |
| Australia | 40.55% | 80.74% | $34,959.77 | 4.95 |
| Austria | 30.34% | 54.85% | $37,476.44 | 5.54 |
| Belgium | 27.97% | 41.63% | $35,012.55 | - |
| Brazil | 63.19% | 87.90% | $11,674.75 | 4.97 |
| Bulgaria | 91.14% | 93.14% | $11,265.37 | 3.91 |
| Canada | 32.05% | 81.27% | $35,644.65 | 5.14 |
| Chile | 58.26% | 93.01% | $15,928.82 | 4.02 |
| Chinese Taipei | 52.95% | 75.41% | - | |
| Colombia | 61.08% | 94.41% | $11,096.56 | 4.56 |
| Costa Rica | 63.76% | 89.18% | $13,107.10 | 6.36 |
| Croatia | 55.57% | 94.51% | $20,338.71 | 4.39 |
| Czech Republic | 63.15% | 82.93% | $22,389.24 | 4.03 |
| Denmark | 11.31% | 33.05% | $36,953.21 | 7.95 |
| Estonia | 68.85% | 81.69% | $23,108.89 | 5.37 |
| Finland | 27.31% | 61.53% | $33,533.66 | 6.39 |
| Germany | 17.15% | 52.13% | $34,155.12 | 4.91 |
| Greece | 17.28% | 44.56% | $24,149.77 | 3.31 |
| Hong Kong | 69.36% | 93.68% | $38,086.50 | 4.07 |
| Hungary | 66.05% | 92.20% | $17,543.21 | 4.98 |
| Iceland | 35.66% | 90.63% | $36,162.02 | 7.25 |
| Indonesia | 94.66% | 94.87% | $7,058.77 | 3.39 |
| Ireland | 32.05% | 66.57% | $38,506.62 | 5.04 |
| Israel | 82.98% | 90.54% | $28,377.45 | 5.74 |
| Italy | 40.44% | 74.43% | $31,768.17 | 4.46 |
| Japan | 79.36% | 52.53% | $31,167.26 | 3.57 |
| Kazakhstan | 99.68% | 99.63% | $20,204.95 | 3.06 |
| Jordan | 77.06% | 87.63% | $9,278.01 | - |
| Korea | 55.98% | 70.74% | $25,180.40 | 4.65 |
| Latvia | 88.71% | 96.27% | $13,099.90 | 5.49 |
| Lithuania | 72.35% | 94.62% | $21,070.60 | 5.21 |
| Luxembourg | 17.18% | 37.73% | $74,278.36 | 4.09 |
| Macao (China) | 76.85% | 81.94% | $79,065.15 | 2.88 |
| Malaysia | 92.21% | 97.63% | $21,210.07 | 5.86 |
| Mexico | 77.78% | 88.70% | $13,444.35 | 4.89 |
| Montenegro | 73.55% | 75.56% | $13,443.99 | - |
| Netherlands | 52.03% | 77.26% | $39,424.83 | 5.20 |
| New Zealand | 57.75% | 97.86% | $27,585.72 | 6.63 |
| Norway | 21.55% | 74.91% | $49,879.99 | 7.10 |
| Peru | 78.47% | 84.15% | $8,880.66 | 3.08 |
| Poland | 81.20% | 94.67% | $15,787.43 | 5.05 |
| Portugal | 35.84% | 84.42% | $23,120.34 | 5.24 |

|  |  |  |  |  |
|---|---|---|---|---|
| Qatar | 85.61% | 93.18% | $120,872.50 | 3.46 |
| Romania | 55.27% | 71.53% | $12,730.27 | 3.24 |
| Russian Federation | 98.72% | 94.56% | $13,457.82 | 3.43 |
| Singapore | 86.45% | 99.10% | $69,612.07 | 3.07 |
| Slovak Republic | 75.02% | 85.42% | $20,612.61 | 4.06 |
| Slovenia | 39.32% | 91.90% | $28,179.14 | 5.61 |
| Spain | 43.01% | 76.44% | $27,740.72 | 4.42 |
| Sweden | 24.51% | 89.89% | $36,285.65 | 7.03 |
| Switzerland | 30.74% | 36.88% | $45,845.67 | 5.11 |
| Thailand | 83.80% | 92.56% | $11,047.45 | 4.30 |
| Tunisia | 68.40% | 88.30% | $8,961.68 | 6.55 |
| Turkey | 62.65% | 85.74% | $16,261.99 | 3.87 |
| United Arab Emirates | 92.81% | 95.80% | $59,688.20 | - |
| United Kingdom | 83.75% | 98.10% | $32,235.57 | 4.70 |
| United States | 57.34% | 94.95% | $43,952.09 | 5.19 |
| Uruguay | 36.11% | 82.81% | $14,856.52 | 2.07 |

Note: Survey weights incorporated.