The Pennsylvania State University

The Graduate School

College of Education

THE GLOBAL EXPANSION OF THE TESTING CULTURE: NATIONAL TESTING POLICIES AND THE RECONSTRUCTION OF EDUCATION

A Dissertation in

Educational Theory and Policy

by

William C. Smith

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2014

The dissertation of William C. Smith was reviewed and approved* by the following:

David P. Baker
Professor of Education and Sociology
Dissertation Advisor
Committee Chair

Katerina Bodovski
Associate Professor of Education

Soo-Yong Byun
Assistant Professor of Education

Molly Martin
Associate Professor of Sociology

Gerald K. LeTendre
Professor of Education
Department Head of Education Policy Studies

*Signatures are on file in the Graduate School.

ABSTRACT

To ensure equal access to high quality education, the global expansion of universal basic education has included accountability measures in the form of academic tests. As the global community adjusts from educational access to educational quality, educators (teachers and principals) are increasingly held accountable for student test scores. Testing for accountability attempts to modify educator behavior by connecting formal or informal, positive or negative consequences to school aggregate test scores. This study examines the association of this international phenomenon with school structures and student outcomes. The investigation begins by exploring the global transformation toward testing for accountability in light of the emerging global culture, identified by World Culture theorists. As testing for accountability emerges as a legitimate global practice questions arise as to how it will shape education and impact students. To address these questions this study maps participants from the 2009 Programme for International Student Assessment into four National Testing Policy (NTP) categories: Formative, Summative, Evaluative, and Formal Sanction/Reward. Using a three level random coefficient Hierarchical Linear Model to capture the nested nature of the data three research questions are addressed: (1) How is national testing policy related to the degree schools incorporate testing into their practices and policies? (2) Does the incorporation of testing into a school structure vary by the school's economic and/or academic composition? And (3) how does the national testing policy and corresponding policy coupling influence student outcomes?

Results suggest that schools in testing for accountability systems are more likely to use student test scores for school accountability purposes. Additionally, within policy differences speak to important equity concerns as less advantaged schools in accountability systems where formal sanctions and/or rewards are applied to schools or educators are more

likely to transfer out low achieving students. Furthermore, initial differences in student achievement across NTP categories are an artifact of schools using their admission and transfer policies to shape their student body. This lack of relationship between NTP and student achievement calls into question whether the substantial costs associated with testing for accountability policies are an efficient use of resources.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I want to thank the numerous individuals who have pushed my thinking, challenged me to work harder, think beyond the norm, and shaped me as an advocate for public education globally.

First, thank you to the two women who most significantly influenced who I am today: my mother, Pamela Smith and my wife, Anna Persson.

Thank you Mom for your endless love and support. You knew the importance of education, encouraged me to do something I loved, and challenged me to make a difference in the world. I am proud to be the first member of our family to complete not only a bachelors' degree but a doctorate.

Thank you Anna for your patience and selflessness. Without you by my side these past 15 years I would have never completed my doctorate. You are amazing. Your smile kept me going through the challenging times and your sage advice reminds me that work needs to be balanced with life. Now that I am finally done being a student, I cannot wait to see what life has in store for us.

Additionally, I want to thank my colleagues at Penn State, especially the preceding cohort which welcomed me to the Education Theory and Policy program and made me feel immediately at home and my cohort whose enthusiasm and activism was inspiring, benefiting everyone in Rackley.

Thank you to those that worked with me on David Baker's research team for teaching me how to collaborate and your persistence in projects that often took years to complete.

Thank you Devin Joshi, my mentor and colleague at the Josef Korbel School of International Studies, for encouraging me to explore the connection between education and

development. Your class has significantly shaped my understanding of education as a public good and drives much of my comparative international research.

Thank you to David Baker, my doctoral advisor and committee chair, for your unique collaboration now and into the future. Your inclusive approach to graduate students, driven research agenda, and inquisitive discussions will be a model for me as I go forward in academia.

Funding for this study was provided through a Thomas J. Alexander Fellowship with the Organization for Economic Co-Operation and Development (OECD). An edited version of the literature review is forthcoming in *Education Policy Analysis Archives* under the title "The Global Transformation toward Testing for Accountability". A modified version of the results and discussion will be published in late 2014/ early 2015 as an *OECD Working Paper* under the title "National Testing Policies, School Practices, and Student Outcomes: An Analysis using data from PISA 2009." A policy brief discussing the results of the first research question will be published in 2015 in the *PISA in Focus* series under the title "The Relationship between National Testing Policies and School Accountability Practices".

CHAPTER ONE: INTRODUCTION

Education is identified globally as both a human right and a primary channel to mobility. With the increasing importance of education globally, international debate has turned from providing educational access to ensuring efficiency and equity in educational outcomes. Accountability systems have been hailed as the answer to issues of efficiency and equity and although few dismiss the need for accountability, how accountability is implemented in education remains a contentious question. The use of 'objective' standardized tests is often proposed as part of the policy solution. Standardized tests provide a means of cross-school, cross-region, and cross-nation analysis. The disaggregation of results by vulnerable groups can reveal equity gaps in education. The use of national testing for school accountability has gained momentum since the passage of the 2001 No Child Left Behind act in the United States. Between 1995 and 2006, the number of countries participating in a national standardized test doubled and by 2006 national tests had been administered in over half of the world's developing countries and 75% of developed countries (Benavot & Tanner, 2007). This expansion has been matched by a qualitative shift in test emphasis and aims, with schools increasingly held accountable for their students test scores. To fully understand this transition it is important to differentiate between traditional high-stakes tests, which place responsibility directly on the student, and testing for accountability, which places responsibility for student performance on educators (teachers and administrators).

Testing for accountability systems are often implemented to ensure efficiency and equity in the delivery of high quality education. Over the past decade research has revealed positive and negative, intentional and unintentional consequences of testing for accountability (de Wolf & Janssens 2007; Hanushek & Raymond 2004; Lee 2008).

Countries that practice testing for accountability have seen marginal gains in student achievement scores. Critics, however, point to evidence that testing for accountability encourages educators and schools to narrow the curriculum, reshape the testing pool, and focus their resources on students most likely to pass the test. Additionally, by identifying low achieving students as potential liabilities that depress the school mean score, testing for accountability fails to decrease equity gaps.

To examine the relationship between testing for accountability, school practices and student outcomes, this study mapped participants from the 2009 Programme for International Student Assessment (PISA) onto four National Testing Policy (NTP) categories: Formative, Summative, Evaluative, and Formal Sanction/Reward (hereafter Form Sanc/Rew). Formative and Summative systems do not provide publically available data and therefore do not fit into traditional definitions of accountability. Formative countries use exams primarily to inform instruction while Summative countries use exams to summarize the student's achievement for the individual student, their parents, and their teachers. Evaluative accountability systems make comparable student test scores public knowledge, providing parents and community members with the necessary information to evaluate school quality on the basis of such scores. The rationale behind Evaluative accountability systems is drawn from economic theory, positing that the disclosure of information about schools create informed consumers, in the form of parents. Additionally, the publication of systematic, comparable information rectifies problems with imperfect information (Hanushek & Raymond 2004) and helps overcome the principal-agent problem (Figlio & Loeb 2011; Woessmann 2007). Form Sanc/Rew accountability systems attach rewards or formal sanctions to the compared outcomes. Form Sanc/Rew accountability functions through a behaviorist model, suggesting individual actions are molded through

incentives and punishments (Hanushek & Raymond 2004) which can dramatically shape individual practice and school structure.

Using a three level random coefficient Hierarchical Linear Model (HLM) to capture the nested nature of the data (i.e. students nested in schools nested in countries), this study looked at the effect on student achievement from a number of different variables including selective admission and transfer policies, publicly posting school level results, and using student test scores to evaluate teachers and principals. Specific research questions addressed in this study included: (1) How is national testing policy related to the degree schools incorporate testing into their practices and policies? (2) Does the incorporation of testing into a school structure vary by the school's economic and/or academic composition? (3) How does the national testing policy and corresponding policy coupling influence student outcomes?

In addition to moving past prior investigations that tend to provide a simplistic, binary accountability variable by clarifying NTP categories based on testing for accountability practices, this study added to the larger debate around accountability in education in three ways.

First, results from previous literature have been narrowly focused on academic achievement scores (de Wolf & Janssens 2007; Lee 2008). This study used the four proposed categories to identify variance in school structures and equity in non-achievement student outcomes. By capturing school accountability practices, such as publicly posting school mean scores, and other related school practices, such as the use of selective admission and transfer policies, this study provides a rich description of accountability at the school level, given the differences in national level policy. In addition the inclusion of student level outcome variables beyond student achievement expand our understanding of

how testing for accountability relates to student well-being and shapes their educational experience.

Second, this study tested the assumption that minimal within country variance is present in the implementation of accountability systems (Woessmann 2007) by identifying schools with relatively loose or tight coupling and using the level of coupling as an explanatory variable in student outcomes. This was done through the creation of a school compliance variable where schools that report practices that align closely with their national testing policy were identified as in compliance and exhibit tight coupling. This initial attempt to capture differences between policy articulation and enactment is unique in the field of educational research. Recognizing differences in policy coupling is important in education as educators are often given significant autonomy. Furthermore, the presence of loose coupling can create within-country variation which may play a significant role in student achievement and non-achievement outcomes.

Third, the preponderance of past studies have focused on single countries, with the vast majority centered on three countries, the U.S., the U.K, and Chile (Figlio & Loeb 2011). Woessmann (2007) suggests that single country studies provide limited insights, as national testing policy is uniform across a single country sample. This study moved beyond single country studies by clustering nations into NTP categories and providing a comparative analysis between national testing policies. The stability of NTP categories given the otherwise heterogeneous nature of countries that constitute each category speaks to the strength of identifying differences through cross-policy comparative analysis.

As testing is legitimated as an essential part of national level education policy and testing for accountability expands to developing countries, questions arise regarding which policy is appropriate to adopt and what are the likely implications of the policy on

educational quality and equity. This study can aid policy makers faced with this dilemma by identifying relationships between national level policy, contextual factors and student outcomes. Formative, Summative, Evaluative, and Form Sanc/Rew testing policies shape schools differently and lead to variations in how schools use achievement data, if they make the data public, the number of assessments students participate in, and how achievement information is used for enrollment and student transfers. Equity gaps in these school structure outcomes between high and low achieving schools and high and low SES schools are also identified. The results provide insight to policy makers, helping them target desired outcomes given their country context. Additionally, patterns which suggest similar school and student outcomes across national testing policies can indicate potential spending inefficiencies. For example, since no significant difference is found in mathematics achievement between countries which provide their school level test scores to the public (Evaluative systems) and those that take accountability one step further by the linking the school level results to formal sanctions and/or rewards (Form Sanc/Rew systems), it would be unwise for countries to waste resources that mandate compliance through punitive measures.

Additionally, since policy makers are cognizant of their national policies and context they can use the results on the importance of coupling to inform their future decisions. If loose coupling is beneficial, given a specific national testing policy, policy makers might consider granting greater local autonomy or using incentives in place of mandates in policy levers. Conversely, if tight coupling is beneficial, policy makers may try to mandate compliance and increase local oversight to ensure policy implementation is standardized across local contexts.

This study proceeds in four chapters. The next chapter (chapter 2) follows the expansion of testing globally, clarifies the aims of testing and introduces the NTP categories. National level examples from Hungary, Mexico, and South Korea are provided to demonstrate the international transformation toward testing for accountability. World Culture theory is explored as one potential framework to understand the adoption of more stringent accountability measures by countries. As testing for accountability becomes increasingly recognized as a scientific practice embedded in the larger dominate culture the adoption of such policies will become normative, with few countries willing to dissent from the trend. The chapter continues by reviewing the potential benefits and consequences that accompany Evaluative and Form Sanc/Rew testing policies. Finally, the chapter presents an open ended question, are we moving toward a normative testing culture. The remainder of the study is a first step in exploring this larger query.

Chapter three introduces the data and methods used in this study. Of particular interest here is the mapping of participants from the 2009 PISA into NTP categories. This categorization scheme, based on policy documents and third party reports and validated by within country experts, is the first of its kind. It moves beyond prior attempts to categorize countries by including a more diverse set of countries (beyond the OECD) and focusing on testing for accountability instead of traditional high-stakes tests. Given the nested nature of the data, students nested in schools nested in countries, a multi-level approach is used with Hierarchical Linear Modeling (HLM) or Hierarchical Generalized Linear Modeling (HGLM) applied as appropriate. The analysis is introduced in two stages with research questions one and two addressed in the first stage which features a two-level model, students nested in countries. The second stage of the analysis addresses research question three and

features a three-level model. Rare in education research this three stage model predicts student outcomes given the NTP and accompanying school practice.

Results are presented in chapter four. The chapter starts with a descriptive overview of included variables by NTP category. Baseline differences are then identified through One-Way ANOVA and chi-square bivariate analyses. To examine whether some nations may be more likely to fall into a specific NTP category, a multinomial regression was conducted predicting NTP from national level variables. Finally, multi-level analyses are completed for each research question with results reported for school accountability practices, school compliance, other related school policies, as well as student academic and non-academic outcomes.

A discussion chapter concludes the study. This chapter is broken down into four main sections. First, the study aims and contribution to the field are reviewed. Second, the main findings are interpreted in light of past research. Third, a large set of policy recommendations are suggested based on the results of this study and the preponderance of past research. Lastly, limitations of this study are identified and directions for future research are put forth.

CHAPTER TWO: LITERATURE REVIEW

With the importance of education increasing globally, international debate has turned from providing educational access to ensuring efficiency and equity in educational outcomes. To certify the available education is of high quality, the global expansion of universal basic education has included accountability measures in the form of academic tests. Between 1995 and 2006, the number of countries implementing standardized national tests more than doubled and by 2006 over half of all developing countries and three-quarters of all developed countries have administered at least one national test (Benavot & Tanner, 2007). This expansion has been matched by a qualitative shift in testing characteristics, increasingly making schools accountable for their student's test scores.

Understanding the transformation from traditional high-stakes exams that place responsibility of test scores on the student to testing for accountability, which places intentional or unintentional positive or negative consequences on educators (teachers and administrators) for their student's performance, is important because different approaches to testing are likely to lead to variant student outcomes (Harris & Herrington, 2006). This chapter investigates this global transformation through an exploration of the components of the emerging global culture, as identified by World Culture theorists. World Culture theory is often criticized by both proponents and opponents as merely a descriptive theory that fails to consider the potential outcomes of the emerging world culture (Carney, Rappleye & Silova, 2012; Schofer et al., 2012), this chapter explains how the self-proclaimed components of world culture support and perpetuate the movement toward testing for accountability.

The chapter starts by using data from international assessments to illustrate the rapid expansion of testing and contrast the current turn toward testing for accountability with more

traditional understandings of high-stakes examinations. This is followed by a historical look at the early adopters of testing for accountability, the United States and the United Kingdom, and the role of New Right ideology in their testing reform. Sections four and five situates the testing for accountability trend within the contemporary "global education reform movement" (see Sahlberg, 2010), introduces World Culture theory, and explain how multiple elements of the emerging world culture legitimate testing for accountability systems. To understand the shift toward greater accountability, National Testing Policy categories are outlined in section six and illustrative national examples from Hungary, Mexico, and South Korea are provided in section seven. Section eight outlines the potential benefits and remaining concerns when schools and systems move toward testing for accountability. Finally, the concluding section asks whether this global transformation is moving the world toward a normative testing culture that has the potential to influence multiple facets of society.

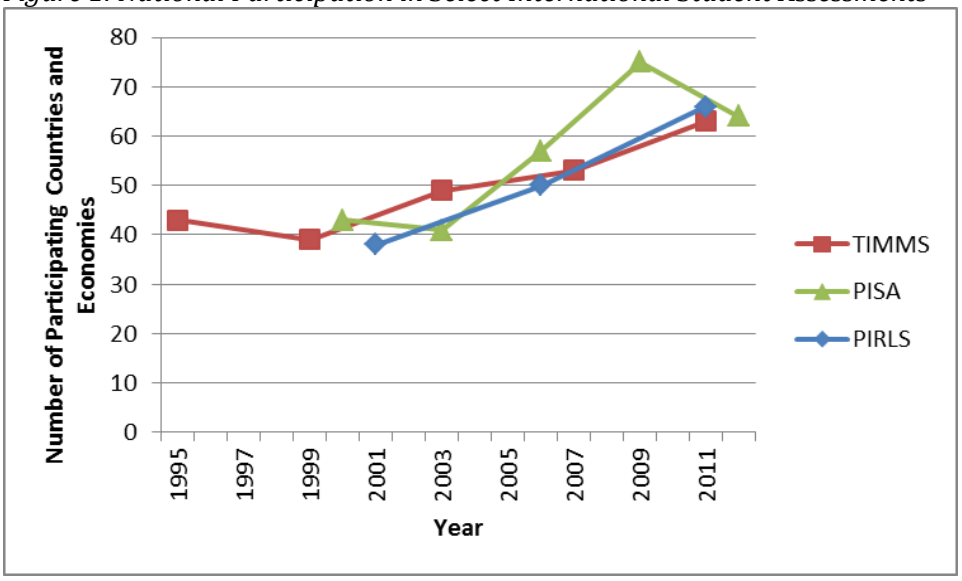
The Expansion of Testing

Testing has long been used to assess student understanding, inform instruction, and identify students for academic advancement. However, the latter half of the 20th century signified a shift in the type of test administered, illustrated by a sharp rise in the use of large scale standardized tests. In investigating the educational systems of 21 industrialized countries between 1974 and 1999, Phelps (2000) found that 18 increased the number of annually administered large scale tests, leading him to conclude that there is a "clear trend towards adding, not dropping testing programs" (p. 19). Since 1980 nearly all European countries have adopted national testing policies (Eurydice, 2009a). This trend is not limited to the industrialized north as educational reformers around the globe insist that "improving national (or state) testing systems is an important, perhaps the key, strategy for improving

educational quality" (Chapman & Snyder, 2000, p. 457). The acceleration in national test policy adoption is perhaps best illustrated in the work of Benavot and Tanner (2007). They found that between the years 1995 and 2006 the number of countries worldwide that participate in an annual national testing program more than doubled from 28 to 67. As of 2006, 81% of developed countries and 51% of developing countries have conducted at least one national test.

Concurrent with the rise in national testing programs is increasing participation in cross-national assessments. The first cross-national assessments were initiated in the 1960s and were originally regionally focused with participation solely from industrialized countries. For example, twelve countries participated in the First International Mathematics Study in 1964 with only Japan and the United States located outside of Europe. However, since the 1990s international assessments have included a diverse array of countries outside the industrialized world as well as provincial economies, such as Shanghai, China and Dubai, United Arab Emirates. As illustrated in Figure 1, this has resulted in a steady increase in the number of participants in the three largest international studies: Trends in International Mathematics and Science Study (TIMSS), Program for International Student Assessment (PISA), and Progress in International Reading Literacy Study (PIRLS). All studies show a roughly 50% increase in participation between 1995 and 2012.

Figure 1: National Participation in Select International Student Assessments



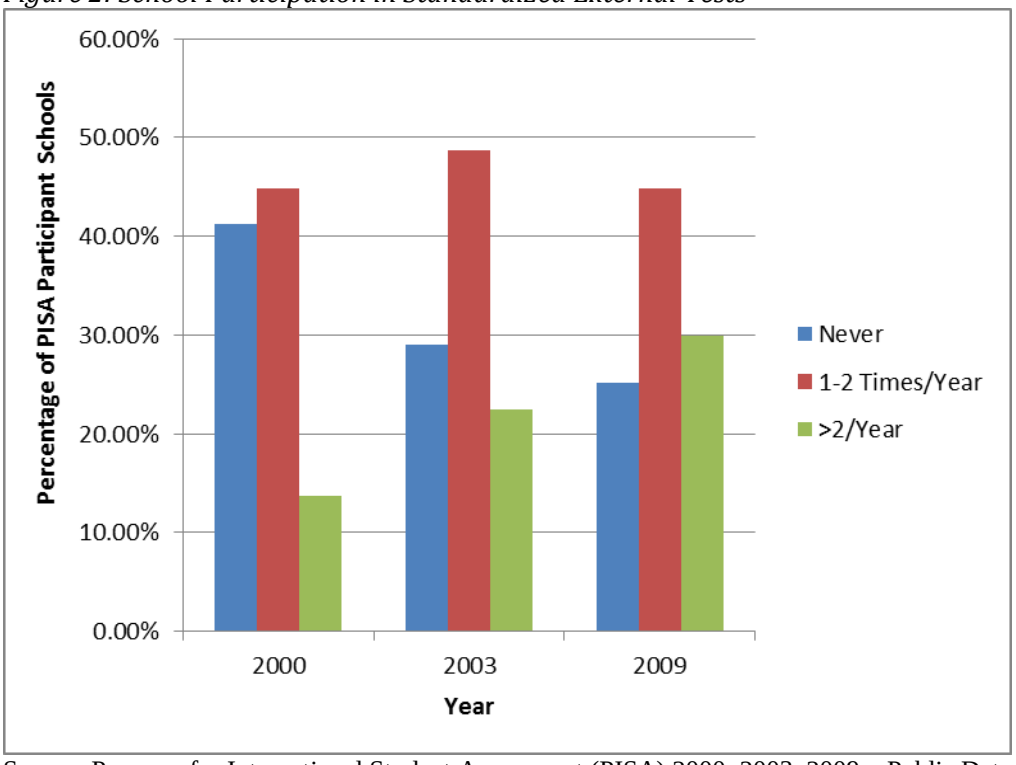
Source: National Center for Education Statistics (https://nces.ed.gov/surveys)

With the support of international agencies, the late 1990s also saw the creation of regional assessments in developing regions. For example the Southern and Eastern African Consortium for Monitoring Educational Quality (SACMEQ) completed their first round of data collection in seven countries in 1999. Since the initial assessment of reading literacy among sixth graders, the number of participant countries has more than doubled, with 16 countries currently partaking in SACMEQ IV, scheduled for completion in 2014. In Latin America, participation in the Latin America Laboratory for Assessment of the Quality of Education (LLECE) has also increased, although at a slower rate, from 13 countries in 1997 to 17 countries in 2006.

At the school or classroom level a shift is also evident in the types of tests students complete, moving away from strictly teacher administered tests to large scale standardized tests. Using data from PISA questionnaires, figure 2 illustrates that the percentage of schools that participate in more than two standardized tests a year have increased from 13.7% to 29.9% over just a ten year span. The number of standardized tests students take

varies greatly across countries with many students exposed to a substantial amount of standardized tests over their compulsory school career. For example, in Denmark students take 36 national tests during their time in primary and lower secondary school, while in China students take up to nine standardized subject tests each year (Schmidt, Houang, & Shakrani, 2009).

Figure 2: School Participation in Standardized External Tests



Source: Program for International Student Assessment (PISA) 2000, 2003, 2009 – Public Data

Turn toward Testing for Accountability

This global transformation is not limited to the number of tests but instead encapsulates a qualitative shift in testing characteristics and aims. 'High-stakes exams' have a long history in many countries. Traditional high stakes exams focus on student knowledge with test outcomes determining student's academic and career trajectory (Eckstein & Noah, 1993). Perhaps the best historical example of a high-stakes test comes from the Chinese Civil Service Exam. Originating in the third century B.C. and formally instituted during the T'ang dynasty (618-907AD), the Chinese Civil Service Exam was based on the Confucian

belief that selection into the ruling class should be based on individual merit (Eckstein & Noah, 1993). The test was originally composed of six distinct examinations but was later narrowed down to one exam, the chin-shih examination, which remained in place until 1904. By the Ming period (1368-1644) the exam was seen as the only legitimate route to government positions as it was "more objective and less open to particularistic influence than the recommendatory system" (Ho, 1964, p. 15). Additionally high-stakes testing for students has been in place in countries throughout Europe, including Iceland, Portugal, and the United Kingdom (UK), since at least the mid-1940s (Eurydice, 2009a).

While much of the early literature muddles high stakes testing by combining effects on students and teachers into one general category (see Chapman & Snyder, 2000; Pari & McEvoy, 2000), the current transformation toward testing for accountability shifts the high stakes from students to educators. Teasing out tests that place high stakes on students from those that place high stakes on educators is important because they provide a different set of motivations and are likely to lead to diverse student outcomes (Harris & Herrington, 2006). The identification of the dominant features of a test, including where the test is administered and to who are the test results focused on, can illuminate these differences. Table 1 identifies tests that are administered within the K12 structure that make educators responsible for their students test scores as *testing for accountability*. Testing for accountability includes the application of formal or informal, positive or negative consequences on educators dependent on their students' performance measures (Figlio & Loeb, 2011). Educational systems that apply testing for accountability are interested in ensuring those that are delegated authority to educate children are "answerable to another level of authority for their prescribed responsibility" (Smeed & Victory, 2010, p. 28), explicitly answering the essential questions linked to accountability: accountability "to

whom" and accountability "for what" (William, 2010). The contentious nature of testing for accountability indicates that answers to these questions remain challenged (Dorn, 1998; Kornhaber, 2004a).

Table 1: Test Foci and Administration

	Focus on Students	**Focus on Teachers/Schools**
Administration Within K12 School Structure	Graduation Exam	Testing for Accountability
Administration Outside K12 School Structure	College Entrance Exam	Teacher Certification Exam

Tests typically associated with high-stakes testing are illustrated in the first column of table 1. Within the K-12 school structure these tests are compulsory for all students (Eurydice, 2009a). High stakes tests that determine the academic trajectory (via tracking) of a student or provides them with access to a subsequent education level can be identified as *testing for advancement.* This differs from *testing for assessment,* low stakes exams which focus on students but are designed to assess their academic progress and direct instruction. *Testing for accreditation* is present when the aim of a test is to provide a credential, identifying an individual as a member of a distinct social group. Teacher certification exams and legal Bar exams are both examples of testing for accreditation. The categorization of testing aims is not mutually exclusive; for example, a high school graduation exam performs the function of both testing for advancement and testing for accreditation.

Holding different actors responsible for student achievement scores has shifted the blame from low performing students to low performing schools (Apple, 1999). When test scores remain aggregated at the individual level, parents often feel a personal interest in improving the quality of education in the classroom, often leading to a more collaborative relationship with the teacher as both the parent and teacher work together to improve student learning. In contrast, the aggregation of results to a classroom or school level allows parents

to place blame squarely on the teacher/school, leading to a more conflictual relationship in which the community questions why the school isn't maximizing the students learning while taking little responsibility on itself. This hostile relationship may become increasingly common as tests for advancement and tests for assessments are increasingly transforming into tests for accountability through school level aggregation. In some countries this evolution is so great that testing has become synonymous with accountability (Froese-German, 2001), suggesting testing no longer has other aims.

The ultimate outcome of testing for accountability reform is largely shaped by educators who play the dual role of policy implementer and student influencer. Teachers, through their position in the classroom hierarchy and student's perception of teacher as legitimate classroom authority, are in a position to significantly influence their student's behavior and cultural understanding (Smith, 2012a). The educator position encompasses both autonomy and obligation. From this position educators have to prioritize often competing demands while continuing practices that are in the best interest of their students (McLaughlin, 1991; Shulman, 1983). The perceived best interests of the student may include adapting classroom routines, structures, or instructional practices to individual student's academic needs. Therefore, unlike student focused testing which affects one student at a time, testing for accountability affects a classroom full of students simultaneously by leveraging behavior change in educators.

Educators as local policy enactors have been identified as "street level bureaucrats"– individuals that have the ultimate responsibility of policy implementation (Shulman, 1983). With autonomy, educators interpret or make sense of the policy, shaping how it is implemented as well as the resulting consequences (Rosen, 2009). Differences in the beliefs and experiences of educators can lead to policy enactment that is substantially different than

the originally articulated policy. Measures of policy coupling have been used to identify the correlation between policy articulation and enactment with low correlation indicating loose coupling and high correlation indicating tight coupling (Datnow & Park, 2009; Shulman, 1983). Although testing for accountability imposes on and restricts educators' professional autonomy (Luna & Turner, 2001), external accountability measures have been successful in altering the in class and administrative decisions of educators (Booher-Jennings, 2005). Of great concern for teachers is their personal and professional survival (Gilles, Cramer & Hwang, 2001) and when faced with testing for accountability this survival instinct is heighted (Nicols & Berliner, 2007), leading "educators who feel oppressed by an ineffective and potentially harmful evaluation system [to] feel justified" (Paris & McEvoy, 2000, p. 150) in altering their behavior to maintain their livelihood. Given the autonomy and authority of educators, the linkage of student success with educator survival has the potential to create large scale changes in the academic quality and success of multiple students simultaneously; making testing for accountability policies substantially different from student focused testing.

Early Adopters of Testing for Accountability

The shift towards testing for accountability has been a relatively recent phenomenon, seen first in the United States and the United Kingdom. The movement towards testing for accountability in the U.S. is rooted in a national history that wants "both a system that rewards merit and a system that generates equality" (Dorn, 2007, p. xiv). Linking tests to student accountability in the U.S. started in the 1960s and by the end of the 1970s many states established a link between test scores and school accountability (Dorn, 2007). The 1970s was characterized by the entrenchment of human capital arguments for education as an investment (Becker, 1962; Schultz, 1961). When combined with the shift in school

funding from primarily local sources to a mix of local and state control, this investment perspective created an atmosphere where "legislators wanted some quid pro quo for spending more on education" (Dorn, 2007, p. 6). The 1970s was dominated by minimum competency standardized tests which mirrored the decline of intelligence testing, the later due to civil rights challenges against using IQ for student placements and an increasing desire to dispel human potential as fixed (Kornhaber, 2004a).

The 1980s saw the rise of the New Right in the U.S. and the UK. Beginning with the governments of President Reagan in the U.S. and Prime Minister Thatcher in the UK (Figlio & Loeb, 2011), New Right calls for improved schooling were articulated "by national policymakers into an umbrella of neoliberal and neoconservative…reforms" (Carl, 1994, p. 315). This ideological shift was largely accepted by the public because of an increasing discontent and distrust of public administrators (Dorn, 2007), due in part to: (1) the economic recession of the 1970s which led to a decline in social services as the 'crisis of the welfare state' led to calls for increased fiscal accountability (Hopmann, 2008), (2) an increasing anxiety about the state of U.S. schools, (3) the inability of desegregation attempts to open up mobility paths for minority students while concurrently threatening the privileged position of the middle class and white parents, and (4) the alienation of working class and minority families from the traditional education experience (Carl, 1994).

The neoliberal push in education from the New Right was dependent on their belief that private schools were more dynamic and innovative than the rigid and bureaucratic public schools, largely because they were situated within a market (Carl, 1994). President Reagan and other New Right supporters used the work of Friedman (1955, 1962) and Chubb and Moe (1990) to justify their position that markets are the solution to a failing school system. Friedman (1962) believed the privatization of education would create a higher

quality, more efficient product while Chubb and Moe (1990) suggested private schools were more efficient due to their organizational structure that provides principals with more autonomy and power. Forcing public schools to compete with private schools would, therefore, result in either an improvement in the product or a closure of poor performing schools. The neoliberal emphasis results in the promotion of individual responsibility amongst self-interested actors, effectively removing any potential societal blame (Hursh, 2007).

The neoconservative call for uniformity and standardization complimented the neoliberal push for market based accountability. Concerned with the relative permissiveness of the 1960s and 1970s, neoconservatives identified the school system as one in crisis (Carl, 1994). To remedy the crisis neoconservatives believed the education system must create uniformity in classroom curriculum and increase enforcement mechanisms (Hill, 2006). In the 1980s, instead of reflecting nostalgically on the past, neoconservatives pressed for "the development of coherent prescriptions for change – usually by hitching the neoconservative cart to the neoliberal horse" (Carl, 1994, p. 300).

New Right Policy Shifts in the U.S.

The racial achievement gap in the U.S. was part of a publicly perceived crisis, increasing concerns about the state of schooling in America during the 1970s and 1980s (Tyack & Cuban, 1995). Measuring the achievement gap required a shift in funding and reporting from inputs of education to outputs – typically measured in test scores (Hanushek & Raymond, 2004; Supovitz, 2009). The failing of the American school system and its inability to reduce the achievement gap were captured in the 1983 report, *A Nation at Risk* (Hopmann, 2008). The report and supporting rhetoric of the New Right helped shape the problem of U.S. education by "implicitly or explicitly attributing responsibility to particular

individuals, institutions, or conditions" (Rosen, 2009, p. 276). The association of a problem with a solution is more easily accepted by the public when it is put in simplistic terms and aligns with already established cultural beliefs (Rosen, 2009). During the 1980s, the simple problem was poor performing schools, shifting the target of accountability from the individual to the schools (Lee, 2008). This "distinct change in direction and philosophy" (Harris & Herrington, 2006, p. 227) resulted in an enormous increase in the number of states with an accountability system from four in 1993 to forty in 2000 (Hanushek & Raymond, 2005). Testing for accountability was also linked to the modern school choice movement, a New Right push to put increased pressure on schools through market based consumer choice. Comparative school level results, produced in testing for accountability systems, would help inform parental decision making. The result was a significant increase in choice options in the U.S. in the 1980s and early 1990s (Eurydice, 2009a; Smith & Rowland, 2014).

Neoliberal ideas have dominated education reform in the U.S. since *A Nation at Risk* (Hursh, 2007). In the 1980s, Texas became the first state to implement testing for accountability (Yarema, 2010). This movement gained prominence nationally when President Bush and state governors met in 1989 to endorse the tying of student test scores to school performance. At approximately the same time the movement to standardize curriculum and instruction was gaining momentum (Hopmann, 2008). Standards were seen as essential for equity, ensuring that everyone was held to the same high expectations (Stotsky 2000). Testing for accountability was encouraged because aligning tests with higher standards would make the tests worth teaching to (Cohen & Ball, 1999; Spalding, 2000; Viadero, 1994), especially if the high standards measure important content (Koretz, 2008). The idea of national standards was embodied by President Clinton's "Goals 2000"

which was supported by the Educate America Act of 1994. Goals 2000 pushed for national standards and implemented voluntary national testing in grades 4, 8, and 12 (Carl, 1994).

In 2001, the reauthorization of the Elementary and Secondary Education Act (ESEA), known as No Child Left Behind (NCLB), became the first national framework linking standards, assessment, and accountability (Datnow & Park, 2009). NCLB linked school performance with student scores on standardized examines and can be understood as "an evolution of previous attempts to use high-stakes tests to improve educational outcomes" (William, 2010, p. 110). Schools were judged on their ability to make adequate yearly progress (AYP) towards 100% student proficiency on achievement tests by 2014. Schools that failed to reach AYP for three consecutive years were subject to corrective action, including potential school closure (Springer, 2008). The emphasis on standardized tests was a boon for the testing industry who recorded a massive increase in test sales in the U.S. from $260 million annually in 1997 to $700 million annually in 2008, a lower bound estimate that does not take into account test support materials and services (Frontline, 2008).

New Right Policy Shifts in the UK

In the 1980s the conservative government of the United Kingdom, under Prime Minister Thatcher, introduced competition between schools to reduce public expenditure on education and encourage the closure of poor performing schools (Stillman & Maychell, 1986). Influenced by New Right ideology, the 1988 Education Reform Act and the 1992 Schools Act established national testing based on a national curriculum for ages 7, 11 and 14, and required local education authorities to produce school level comparable examination results, known as league tables (Edwards & Whitty, 1992; Teelken, 1999; West & Pennell, 2000). Since 1992, league tables have been published annually by the daily national press in England and Wales with public and private schools posted concurrently after 1993 (West &

Pennell, 2000). The 1988 Education Reform Act also established the condition for education as a market by opening enrollment and linking school funding to pupil enrollment (Edwards & Whitty, 1992). The 1991 Parent Charter enhanced the choice environment by emphasizing the rights of parents as active choosers in their child's education, providing means for parents to evaluate the schools based on league tables and relocate their children if necessary to higher performing schools (Teelken, 1999).

Global Transformation toward Testing for Accountability

Although there are some signs that early adopters (i.e. the UK and U.S.) are taking marginal steps away from holding schools accountable this has not slowed the global transformation toward testing for accountability. In the UK, regional autonomy has led Scotland to scrap its testing program (Volante, 2007) and England to eliminate tests for 14 year olds in 2008 (Eurydice, 2009a), however, league tables and between school competition still dominate UK education policy. In the U.S., congress has failed to reauthorize ESEA and provide an alternative to NCLB. With congress deadlocked, President Obama has pushed through his seminal policy, Race to the Top, and provided waivers to states to circumvent some of the requirements put forth by NCLB, namely the arbitrary 2014 deadline for 100% proficiency. Both policies, however, continue the NCLB emphasis of basing school evaluations on comparable school level data which is available to the public and tying test scores to teacher livelihood through the implementation of 'pay for performance' schemes (Dillon, 2011; McNeil & Klein, 2011; Smith & Rowland, 2014).

Regardless of perceived steps away from testing for accountability by the U.S. and UK, the global expansion continues full speed as countries follow the early examples of the U.S. and UK and engage in the "ubiquitous adoption of accountability policies" (Hanushek & Raymond, 2004, p. 407). Testing for accountability is viewed as a common solution to

education problems around the world and is part of the "global education reform movement" (Sahlberg, 2010, p. 47), as illustrated in Butland (2008), Lemke et al. (2004), Teelken (1999) and Figlio and Loeb (2011). Furthermore, accountability mechanisms that leave control to regional or national authorities are substantially similar across countries (Macnab, 2004). The adoption to testing for accountability by diverse countries around the world (see section 7 below for examples) suggests that "the development and implementation of accountability systems has been one of the most powerful, perhaps the most powerful, trend in educational policy in the last 20 years" (Volante, 2007, p. 4).

In many countries donor agencies play a significant role in shaping national policy. Increasingly, multilateral organizations and international finance institutions reinforce testing for accountability by linking loan conditions to assessment infrastructure and policy (Kamens & McNeely, 2010). Of note is the World Bank's movement toward supporting a testing culture. In their content analysis of the World Bank's Education Sector Strategy 2020, released in 2011, Joshi and Smith (2012) find a nearly 100% increase in terms associated with the testing culture from the prior 1999 Sector Strategy paper. This included an astonishing increase in mentions of "accountability" from twice in the 1999 strategy to 32 times in the latest release. This led the authors to question whether a World Bank focus on "test-based education may be crowding out emphasis on well-rounded skills, personality development, and critical thinking that might come from a more problem-solving or dialogic approach to education" (p. 192). The recent establishment of the World Bank's SABER tool provides additional evidence of a global trend towards increased accountability. SABER is a voluntary tool for nations to evaluate the effectiveness of their education system. It is strongly encouraged by the World Bank and includes provisions for national and international assessments. Countries which do not implement a national testing policy that

ensures accountability or fail to participate in international assessments, such as PISA, receive lower grades (Bruns, Filmer, & Patrinos, 2011). Although voluntary in nature, the normative pressure placed on national leaders by SABER pushes countries toward adopting testing for accountability policies.

Explaining the Turn towards Testing for Accountability

World Culture theory suggests that the global acceptance of testing for accountability can be understood as part of a larger cultural and collective process based around shared global values and ideas of legitimacy (Meyer, 1977). World Culture theory is one strand of a larger theoretical framework known as Neo-institutionalism, Sociological Institutionalism, or World Society theory (Schofer et al., 2012; Wiseman, Astiz & Baker, 2013). Neo-institutionalism sees "social action as deriving from culture, knowledge, and authority rooted in global institutions and structures" (Schofer et al., 2012, p. 57). Institutions, such as the family, religion, and education, help construct culture by expanding social roles and legitimating action and knowledge (Baker et al., 2006; Meyer, 1977). For example, "modern educational systems formally reconstruct, reorganize, and expand the socially defined categories of personnel and of knowledge in society" (Meyer, 1977, p. 72). The cultural products of institutions are therefore shaped by the institution, strengthening and reinforcing its authority (Baker et al., 2006). Institutions influence actors at the local, regional, national, and international level through molding the culture they are embedded in (Schofer et al., 2012).

Individual actors recognize what is appropriate behavior within a given culture by the normative scripts or cultural models associated with each social role (Schofer et al., 2012). Scripts guide actors, telling them how to feel or act in the world (Baker, 2014; Baker et al., 2006). As within any culture, deviation from the script is subject to public scrutiny.

The socialization process through which institutions create socially accepted scripts and culturally enforced role compliance often result in the internalization of acceptable and logical ways to engage in the surrounding social environment (Baker et al., 2006). The result is behavior that is taken for granted or understood as common sense and, therefore, beyond question. Essentially, culture has shaped behavior by identifying some actors and actions as legitimate while dismissing others (Jepperson, 2002).

Cultural institutions spur a process of cultural alignment across populations known as isomorphism (Wiseman, Astiz & Baker, 2013). Education, as an institution, is an interesting example of isomorphism. From a neo-institutional perspective, the presence of similar education models globally are the result of shared meanings and values that identify appropriate rules and routines (Baker, 2014; Wiseman, Pilton, & Lowe, 2010). Schools and education systems then "become isomorphic with the institutional environment in order to achieve legitimacy and ensure their survival" (Booher-Jennings, 2005, p. 234).

Components of the World Culture

As an increasingly normative policy lever that provides legitimacy to countries that practice it and reconstructs the notion of education, teachers, and students, testing for accountability is one of the components of the dominant world culture. Embedded within the testing for accountability movement is the taken for granted assumption in neoliberalism and the power of competition to produce quality. Kamens & McNeely (2010) suggest that the growth in assessment internationally represents a move towards a world educational ideology that consists of unfailing faith in science as the path for legitimate knowledge and a belief that organizations can be managed to produce a desired outcome. As illustrated by policy practices in the World Bank, there is a growing international consensus that participation in international testing and the use of a national assessment system are essential

in a legitimate education system. Additionally, the number of countries involved in testing for accountability is likely to increase as more policymakers are "doing what is expected of them by their individual and institutional peers" (Wiseman, 2010, p. 2).

Multiple elements of the emerging world culture provide justification for testing for accountability systems, including: the expansion of western models, an emphasis on academic intelligence, faith in science as the rational path to truth, and the decentralization of authority to the local level.

Western Model

The world culture is western in origin, shaped through the western universities charter to produce legitimate knowledge and spread through the rapid expansion of educational attainment known as the education revolution (Baker, 2014; Ramirez, 2003). As an advanced schooled society, many American ideas of education can foreshadow global outcomes (Baker, 2014). Noticeably the western model preaches education for human development and education as a human right (Baker, 2014; Kamens & McNeely, 2010). Situated within this model is the increasing importance of education in later life outcomes. The position of schooling as both a private and public good encourages countries to implement mandatory policies requiring all children attend (Baker, 2014). Since all students have the ability to achieve, stratified schooling is deemed unjust and decisions regarding equal rights to an education no longer center on access but quality. When combined with a western view of education as "a 'technical' science that can be studied, rationalized, and quantified" (Wiseman, 2010, p. 18), it is not surprising that the right to a high quality education leads to a push toward measurable indicators that can be used for accountability purposes (Kamens & McNeely, 2010).

Academic Intelligence

The education revolution has reinforced the cognitive and scientific dimensions of legitimate knowledge (Baker, 2014). Understood collectively as academic knowledge, this legitimate knowledge emphasizes meta-cognitive skills and the value of empirical evidence (Wiseman, 2010). Subjects that encourage this type of knowledge, namely mathematics and science, are considered valuable as demonstrated by an increasing use of mathematics achievement scores in public comparisons of schools and countries (Baker, 2014). The importance of science and mathematics, as central to academic intelligence, is reflected in the work of Kamens, Meyer, and Benavot (1996) who found these subjects were no longer restricted to specialist knowledge but were now available to and intended for all students. Additionally, less academic subjects, such as the visual arts, have been dismissed for subjects that produce academic intelligence (Baker, 2014).

Faith in Science

Science production measured by the number of scientists, scientific publications, scientific training, and the number of countries with a national science program continues to increase globally (Baker, 2014). The swelling of science production is often called for by policymakers and practitioners who believe science to be an objective arbiter of truth (Rosen, 2009). Critics of the unquestionable faith in science recognize that "as long as the public maintains this irrefutable objectivity of statistics, a graph here and a chart there can leverage support for provincial reforms that could never survive nuanced deliberation" (Robertson, 1999, p. 715). Sciences taken for granted position increases the value placed on education that uses test scores to objectively and accurately measure student knowledge (Paris & McEvoy, 2000). When international test scores are reported their "seemingly authoritative measure of students' skills and abilities" (Cohen & Rosenberg, 1977, p. 128)

prompt nations to respond through the development of appropriate educational reform (Drori et al., 2003).

Decentralization

Decentralization has become a widespread institutional model as "a significant set of nations have responded to the legitimizing global forces within a multinational economy and world institutional system by adopting decentralization" (Astiz, Wiseman, Baker, 2002, p. 86). In Europe the increasing devolution of responsibility to the local level has been met with increased curricular and evaluative control at the regional or national level (Eurydice, 2009a). Testing for accountability is likely in decentralized systems because external exams are required to ensure education quality across diverse communities, where local control often leads to information asymmetry (Woessman, 2004, 2007). From this perspective, "statistical accountability systems" are seen as "one way to resolve the dilemma between granting autonomy and authority to educators and keeping them under some political control" (Dorn, 2007, p. 13).

Neoliberalism

Neoliberalism, as an economic system, has spread to nearly every country in the world (Friedman, 1999). Neoliberalism promotes private property, open markets, and free trade on the basis of three core assumptions: (1) consumers have access to accurate market information, (2) consumers act as self-interested profit maximizers, and (3) private provision and competition will be more efficient than public control (Harvey, 2005; Jolly, 2003). The diffusion of neoliberalism as a legitimate economic approach occurred once it emphasized the general benefits of competition, embraced the role of actor as central to global institutions, and was viewed as a legitimate system by international institutions (Schofer et al., 2012).

As a dominant approach to education policymaking, neoliberal practices are often adopted by countries seeking legitimacy (Wiseman, 2010). The implementation of neoliberal policy reinforces cultural faith in its promise of increased performance with improved efficiency, solidifying the position of neoliberalism as a 'common sense' approach that cannot be questioned (Apple, 1999; Hursh, 2007; Rosen, 2009). Treating education as a market invites the invisible hand of the market to improve the quality of schools through between school competition for students (Apple, 1999; Levin, 1992). For neoliberal policymakers academic gains must be balanced with financial costs (Wiseman, 2010). Public investment in education must be justified (Smeed & Victory, 2010): "the public has a right to expect that its resources are being used responsibly and that the public institutions are accountable for caretaking the public trust" (Supovitz, 2009, p. 215). Test results, therefore, provide an easily measured indicator of quality to ensure the public investment in education is being used efficiently and effectively.

Test Results as Information

The publication of standardized test results provide parents, acting in the role of consumers, with easily interpretable market information allowing them to put more pressure on schools to respond to consumer demand (Ball, 1993; Woessman, 2007). The ability of parents to act rationally as self-interested consumers is a prerequisite to an effective market (Apple, 1999). The use of test results as information helps overcome the principal-agent problem, recognized by many neoliberal economists (Figlio & Loeb, 2011). In situations where a principal (i.e. parents) hires an agent (i.e. educators) to perform a service, if the interests of the principal and the agent do not align, the service may not be performed efficiently. Student test scores provide a common metric of measurement for evaluation, informing the principal of the agent's real performance and providing evidence to ensure

that the needs of the principal are being met (Woessman, 2004). Test results can also be conceptualized as quality indicators which can help direct resources (Joshi & Smith, 2012) and aid the government in targeting inefficient schools to be shut down (Lincove, 2009). The use of test scores in this manner is acceptable as they are often considered the only legitimate measure of quality in education. With this authority, the implementation of testing is increasingly seen as an end in education, instead of a means to improve student understanding (Booher-Jennings, 2005).

Test scores are used to spur parent involvement in the education process (Smith & Rowland, 2014). The UK used their league tables to employ parents to make market choices, with the government telling parents that "you should get all the information you need to keep track of your child's progress, to find out how the school is being run, and to compare all schools" (Department of Education, 1994, p. 3). While there is some evidence to support the use of this information in the parent's decision-making regarding their child's school attendance (Teelken, 1999), others find parents pay little attention to the publication of data (de Wolf & Janssens, 2007). Janssens and Visscher (2004) suggest that parents pay little attention to this information because they do not have access to the information, they lack the capacity to understand the information, they have limited real choice between schools, and the information does not reflect the factors parents base their decision on.

National Testing Policy Categories

As the expansion of world culture legitimates and institutes testing for accountability as a taken for granted education policy it is important to remember that variation in national policy remains. Examining national policy is important because "national ministries of education typically act as agents imposing this activity [testing] on schools and education systems" (Kamens & McNeely, 2010, p. 6). Differences in national testing policy can best

be seen on a rough continuum based on the presence and intensity of testing for accountability (see table 2). This categorization strategy is similar to the scheme of Eurydice (2009a) and the stated purposes of education policymakers often transcend policy categories (Eurydice, 2009a).

Table 2: National Testing Policy Categories

Testing for Assessment	Testing for Assessment, Advancement, or Accredidation	Testing for Accountability	
Formative	**Summative**	**Evaluative**	**Formal Sanction/Reward**
The use of national or regional examinations as diagnostic tools that are used internally by schools to inform instruction.	The use of national or regional examinations as a tool that summarizes student learning and is shared with parents; when disseminated is done so at the national or regional level.	The use of national or regional examinations as a tool that summarizes student learning and is disseminated at the school level to allow for between school comparisons.	The use of national or regional examinations as a tool that summarizes student learning, is disseminated at the school level, with school/class level results used to apply rewards or sanctions.

At the far left of the continuum are *Formative testing policies* which use tests for assessing the progress of students. In this system tests are ongoing, informing teacher instruction through direct feedback (Eurydice, 2009a). Formative policies are often professed by policymakers; however, more often than not this gesture is simply policy rhetoric (Irons & Harris, 2006). *Summative testing policies* use tests to assess, provide accreditation, or direct student's educational advancement. They summarize how well an individual is doing at a given point in time (Eurydice, 2009a; Nitko & Brookhart, 2006) and when scores are disseminated they are done so at the national or regional level. Traditional high-stakes student tests that do not aggregate results at the school level and tests that

disaggregate scores by ethic, economic, or regional subgroup, but not school, fall into this policy category.

Evaluative and Form Sanc/Rew testing policies both apply testing for accountability. In *Evaluative testing policies* test scores aggregated at the school level are used by the public to compare schools and evaluate school quality. In this economically based model, "responsibility is devolved to the individual consumer and the aggregate of consumer choices provides the discipline, of accountability and demand, that the producer cannot escape from" (Ball, 1995, p. 69). Informal consequences in this system result from a consumer choice market mechanism and public stigmatization through "naming and shaming" (de Wolf & Janssens, 2007) or the "scarlet letter" effect (Harris & Herrington, 2006).

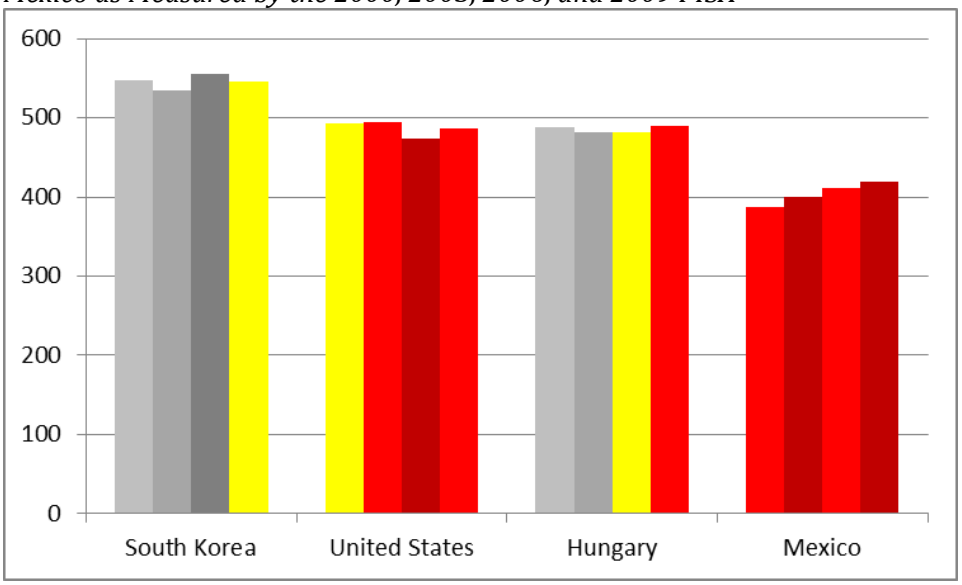
In *Formal Sanction/Reward testing policies* school level aggregate data is used to apply consequences through the application of rewards or formal sanctions. The primary difference between Evaluative and Form Sanc/Rew testing policies can be found in the consequences. Implicit consequences through public pressure characterize Evaluative systems while explicit consequences through formal channels characterize Form Sanc/Rew systems. Form Sanc/Rew systems function through a behaviorist model which suggests individual action can be molded through incentives (Hanushek & Raymond, 2004). Form Sanc/Rew testing policy responds to reformers that believe serious consequences are necessary to transform the education system and has both compliance and avoidance costs, making it expensive to implement (McDonnell & Elmore, 1987).

Examples of National Turns toward Testing for Accountability

To illustrate the turn toward testing for accountability, three national examples are provided from countries that currently employ a Form Sanc/Rew testing policy. As

illustrated in figure 3 the movement toward a Form Sanc/Rew policy is not dependent on the relative achievement of a nation's education system. Similar to table 2, figure 3 uses yellow to indicate an Evaluative testing policy and red to designate a Form Sanc/Rew testing policy. Looking at mathematics test scores across four rounds of PISA (2000, 2003, 2006, and 2009), no evidence of increasing achievement is observed as countries turn toward more intense measures of accountability.

Figure 3: Country Mean Mathematics Scores for South Korea, the United States, Hungary, and Mexico as Measured by the 2000, 2003, 2006, and 2009 PISA



Hungary

Hungary is a country that has traditionally been without a national testing system but has seen dramatic changes over the past two decades (Eurydice, 2009b). After spending 33 years as part of the Soviet bloc, education policies in the 1980s and 1990s provided a lot of freedom, if not a lot of guidance. The 1985 Education Act abolished the previous inspection system but failed to replace it with a viable way to measure education quality. The rapid democratization and decentralization of Hungary following the collapse of the Soviet Union led it to have one of the most decentralized education systems in Europe (Eurydice, 2009a,

2009b). During the 1990s, in a shift that was partially fueled by national performance on international assessments, it became clear that without a national test and with dissention in defining education quality, assessment reform was needed. Starting in 2001, 'monitoring surveys' were implemented yearly and the National Assessment of Basic Competencies (NABC) was established (Eurydice, 2009a, 2009b). Originally testing 5th and 9th graders, the NABC quickly expanded to include 4th, 6th, and 8th grade. The early goal of the test was to develop a within school evaluation culture and allow schools to compare their results to nationally aggregated sub-groups. However, since 2006 results have been disseminated to the general public, providing the school's clients (parents) with information on school effectiveness (Eurydice, 2009a). This Evaluative policy shifted quickly to a Form Sanc/Rew policy when in 2008 schools were mandated to incorporate test scores into internal quality reports. Low achieving schools were then required to use this report to prepare and implement an action plan for remediation (Eurydice, 2009b). As illustrated in Figure 3, during this period of rapid transition toward a Form Sanc/Rew testing system Hungary saw a less than 0.5% improvement in PISA mathematics scores.

Mexico

Since 1993 Mexico has applied a Form Sanc/Rew policy by tying student test scores to explicit rewards for teachers. The Teacher Career Program links the results of the Instrument for Testing New Secondary School Pupils (IDANIS) to teacher bonuses (Ferrer, 2006). Originally established in 1993, the program provides rewards of up to 197% of teacher's base wage. Teachers are evaluated on a myriad of outcomes with IDANIS results representing on average 20% of the teachers overall score. Due to decentralization practices variance is found between regions, however, by 2000 nearly all of Mexico participated in publishing school level data (Ferrer, 2006; Hagerstrom, 2006). In 2001 the quality schools

program (PEC) was established as a voluntary program in which schools could apply for a competitive grant by submitting a five year improvement plan. Increasing in popularity, there are concerns about the financial feasibility of the program (Hagerstrom, 2006). National mean scores in mathematics over the past decade have increased by roughly 8% (see figure 3), however, considering the early establishment of a Form Sanc/Rew system more information is needed to see whether the improvement can be partially attributed to the systems presence, duration, or other unrelated factors.

South Korea

In 1991 South Korea decentralized their education system. Since that time the Korean Institute for Curriculum and Evaluation (KICE) has been responsible for the administration of national assessments. The National Assessment of Educational Achievement (NAEA) is a criterion reference test administered in the 6^{th}, 9^{th}, and 10^{th} grade focusing on Korean, mathematics, science, social studies, and English. Results of the NAEA have traditionally been aggregated and disseminated at the national level. In 2007, plans were unveiled to publish results at various levels, including the school. President Lee Myung-bak declared that moving to an Evaluative policy was essential for school choice to work. Teachers and teacher unions opposed the move on the grounds that Korea was already a high scoring country and national policy should be focusing on 'creative' skills, and as a result school level publication was delayed until 2011 (Schmidt, Houang, & Shakrani, 2009). In 2011, the government began to target low performing schools known as 'creative management schools that pursue academic ability enhancement' (Kim et al., 2010) and "similar to the provisions in the No Child Left Behind Act in the U.S., Korea's core plan is to provide additional support to schools with lots of children who are underperforming, but only for a specific period of time" (Schmidt, Houang, & Shakrani, 2009, p. 56).

Although the policy time horizon is still under development, the end goal is a complete transition to a Form Sanc/Rew policy through the application of sanctions on schools that fail to improve. From 2000 to 2009, as Korea leadership pushed toward punitive policies national mathematics scores on PISA saw essentially no change. However, in the years directly before and after the President's announced change in school test score reporting, the national mean score dropped by approximately 2% (see figure 3).

Implications of Testing for Accountability

With more countries transitioning toward testing for accountability systems it is important to move beyond describing testing as part of the world culture to investigate the substantive outcomes of this global trend (Schofer et al., 2012). Research on the effects of testing for accountability on student outcomes have been mixed and dominated by studies from the U.S., UK and Chile (Figlio & Loeb, 2011). Proponents of linking test scores to market mechanisms and educator rewards and sanctions hypothesize that increased pressure will lead to greater efficiency and quality, increasing student performance (Lavy, 2007) while opponents are concerned with its potential to change educator behavior and influence equity outcomes.

Potential Benefits

Two formal meta-analyses have been conducted on research studies from the United States to estimate the effect of accountability systems on student test scores, with both finding marginal, positive gains. Belfield and Levin (2002) in a meta-analysis of 25 studies that examined the link between competitive pressure and educational outcomes in the U.S., found a modest effect on scores with a 1 standard deviation increase in between school competition associated with a 0.1 standard deviation increase in test score. In a more recent examination of 14 studies exploring the effects of test-driven external accountability

systems, Lee (2008) found small positive effects when mathematics and reading scores were averaged but no effect on the racial achievement gap. Lee concluded that the marginal mean effect size for school accountability (M=0.27) "does not lend strong support for claims for school accountability" (p. 616).

Heterogeneity in the effect of testing for accountability is observed across ethnic group, student ability, subject, and type of accountability. The effect of implementing testing for accountability is inconsistent across studies (Figlio & Loeb, 2011) with some finding it disproportionately disadvantages ethnic minorities (Hanusheck & Raymond, 2005), while others suggest it closes the Hispanic-white achievement gap while increasing the black-white gap (Hanushek & Raymond, 2004), or that it has unilateral benefits for minorities (Carnoy & Loeb, 2003). Using quantile regression to conduct a micro-level analysis across three international datasets, Woessman (2004) concluded that the effect of external exams on student achievement also indicate a relative advantage for higher ability students. Heterogeneity is also present across subjects as research using a multitude of comparison strategies find a positive effect of testing for accountability on mathematics test scores with a weaker or non-significant effect found on reading scores (Cronin et al., 2005; Dee & Jacob, 2009; Figlio & Loeb, 2011; Lee, 2008; Wong, Cook, & Steiner, 2009). A similar trend is found when persistence in achievement gains are explored with mathematics achievement effects persisting one to two years following student transition from a low performing elementary school on the verge of sanctions to a higher performing middle school (Chiang, 2009). Cronin et al. (2005) suggests that differences between subjects may be due to the dependency of mathematics understanding on classroom instruction, while reading is relatively more easily affected by parental involvement.

The type of testing for accountability applied can also lead to divergent results. Studies suggest that Evaluative policies, in which schools are compared through the publication of results, has a positive effect on student test scores, although the "practical significance of this gain is negligible" (Springer, 2008, p. 5). Additionally, in Form Sanc/Rew systems, where explicit consequences are present, student achievement is higher (Dee & Jacob, 2009). When Evaluative and Form Sanc/Rew systems are compared, the relative advantage of Form Sanc/Rew policy outweighs the market pressure of Evaluative systems (Bishop et al., 2001; Hanushek & Raymond, 2005). In comparing eighth grade student test scores before and after the implementation of a Form Sanc/Rew system, Hanushek and Raymond (2005) find a 3.2 scale point increase in scores. As the publication of results was not a significant factor in test scores, the authors concluded that it was the explicit consequences that led to the increase in scores. The gains in test results in testing for accountability system may be partially attributed to the higher expectations of all student groups, including students with disabilities, in the tested subjects (Ysseldyke, Dennison, & Nelson, 2004).

Less studied have been the expected outcomes involving teacher preparation and the shifting of resources to in-need schools. Earl and Torrance (2000) show that in testing for accountability systems teachers participate in more professional development within the tested subjects. This targeted professional development is often complimented with more instructional time for tested subjects (Rouse et al., 2007). The shift in resources to low achieving schools has not been supported by past research, which shows that resources actually flow in the opposite direction, to high performing schools (Chapman & Snyder, 2000). Within schools, resources are shifted from one set of students to another or between subjects depending on the accountability incentive structure (Reback, 2008).

Concerns

Testing for accountability systems provide incentives for actors to game or manipulate the system in order to increase the likelihood of high test scores. The structural and behavioral changes evident in testing for accountability systems result from educators attempting to ensure their survival by focusing their attention and resources on improving test scores, often to the detriment of a well-rounded education. This response of educators is essential for their survival and expected; as Campbell's Law states "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Sahlberg, 2010, p. 52). In testing for accountability the most egregious example of gaming the system involves outright cheating, with educators manipulating test results or completing student examinations (Gipps, 2003; Jacob & Levitt, 2003; Lashway, 2001). More common examples of gaming the system include: shaping the testing pool, narrowing the curriculum, teaching to the test, and teaching to the "bubble". These will be explained briefly below.

Shaping the Testing Pool

Shaping the testing pool changes the group of students subject to testing, generally excluding low performing students to increase the average test score (Hanushek & Raymond, 2004). Low achievers threaten educator survival and in an environment where schools compete in a marketplace these students are identified as poor investments or liabilities (Froese-German, 2001). Shaping the pool can be accomplished by increasing student repetition (Hursh, 2005; Kornhaber, 2004a; Kornhaber, 2004b). Haney (2000) found that students in Texas were more likely to be retained the year prior to the exam and that the increase in grade repetition was disproportionate for minority students.

Reclassifying students into special education is another practice that reduces the number of students that count towards the school test score average (Cullen & Reback, 2006; Haney, 2000; Jacob, 2005). Internationally, special education students are often excluded from aggregate school results (Booher-Jennings, 2005; Eurydice, 2009a). This exclusion is compounded in the U.S. were financial incentives encourage special education classification by providing more money per special education student (Figlio & Loeb, 2011). Once reclassified, special education students are often forgotten about; as one special educator from Texas put it "they [other teachers] actually tell me that since the kids aren't accountable [counted in school aggregate] they're not worrying about them" (Booher-Jennings, 2005, p. 248).

Shaping also occurs through permanent exclusion or expulsion of low achieving students. In New York, schools altered the testing pool by categorizing low achieving students as transferred or working on their GED (Lewin & Medina, 2003). In the UK, Gillborn (1996) found that the permanent exclusion rate in schools increased 300% in the three years following the implementation of league tables. When government officials were asked to explain the dramatic increase, 8% contributed it to increased behavioral issues while 43% linked it to increased between school competition. Figlio (2006) investigated disciplinary records across multiple school districts in the four years surrounding implementation of Florida's accountability system. Focusing on the over 40,000 incidents where at least two students were suspended for the same incident he found that harsher punishments were doled out to the lower performing student and that the difference in punishment increased during the testing season and among testing grades, a pattern not present prior to the accountability system. Figlio concluded that schools use discipline policies to reshape the testing pool by removing low performing students during testing

periods through longer suspensions. When given freedom in admission practices some schools participate in 'cream skimming' by selecting some students while selecting out others (Gerwitz et al., 1995; West, Pennell & Noden, 1998). If combined with movements for privatization, testing for accountability can be more problematic as shaping the testing pool is amplified by private school selection bias (Joshi & Smith, 2012b). As a result low achieving students may not be welcome in either private or public school, resulting in a potentially bifurcated system where lower achieving student are relegated to the few schools that do not practice 'cream skimming'.

Narrowing Curriculum

Testing for accountability spotlights valued knowledge through the inclusion and exclusion of subjects in tests. The limitation of outcomes to easily measured and quantified indictors diverts attention from other, harder to measure goals of education, such as citizenship and social skills (Ben Jaafar & Anderson, 2007; Figlio & Loeb, 2011). Most nations emphasize tests in language arts and mathematics at the elementary and lower secondary level (Eurydice, 2009a). Instructional time is then transferred from untested subjects to mathematics and language arts or basic test preparation (McNeil & Valenzuela, 2001). Rentner et al. (2006) found that over 70% of school districts in the U.S. reallocated instructional time towards mathematics and reading and away from other subjects in response to testing policies. An increase in instructional time, however, does not necessarily indicate improved instruction, as shifts in educator motivation can lead to "more superficial adjustments in content coverage and test preparation activities rather than promoting deeper improvements in instructional practice" (Supovitz, 2009, p. 211, see also Falk, 1996 and McNeil, 2000). Furthermore, the movement away from a well-balanced educational experience focused on depth of understanding is more pronounced in high-minority and

economically disadvantaged schools (Froese-German, 2001; McNeil & Valenzuela, 2001; von-Zastrow & Janc, 2004).

Teaching to the Test

In systems where educator sanctions and rewards are explicitly linked to student test scores teachers have incentives to teach the skills and knowledge that will be included on the test (Hursh, 2007). Teaching to the test involves both attempts to prepare for the content and form of the test as well as mimic the test atmosphere (Au, 2007; Cuban, 2007; Nicols & Berliner, 2007). In preparation for standardized tests, some students are given similarly formatted tests with commonly tested vocabulary for practice (Yarema, 2010). With test gains superseding other intellectual goals of education, teachers may emphasize correct answers over understanding; as one sixth grade teacher for Texas recognized "I have to teach to the test…I need them [students] to know how to solve problems, but not always why and what makes a correct answer" (Yarema, 2010, p. 11-12). In this vein, test preparation often includes explicit instruction (i.e. this is what you need to know for the test) (Certo, 2006). Teachers may also replicate the testing environment to ensure students are familiar with the feel and procedures of the test. For instance, mathematics teachers may have students work independently in rows to mimic the individual tasks required during testing (Yarema, 2010).

Teaching to the "Bubble"

In a testing for accountability system students on the verge of passing the test or on the "bubble" have the ability to greatly impact aggregate school scores and therefore demand the attention of educators. When education is treated as a market, student ability is akin to student value (Hursh, 2007; Teelken, 1999). In such a system "it becomes rational to leave the lowest performing students behind" (Hursh, 2007, p. 506). Teacher preparation

and school practices divert attention and resources away from 'remedial students' by clearly identifying 'bubble kids' or those in the 'strike zone' (Booher-Jennings, 2005; Gillborn & Youdell, 2000; Lipman, 2004; Reback, 2008). A Texas teacher best exemplifies this trend: "I definitely focus more attention on the bubble kids…I feel like we might as well focus on the ones there's hope for" (Booher-Jennings, 2005, p. 242). In this instance the classification and corresponding treatment of students is especially problematic as students are classified into ability groups at the beginning of formal school and little movement is possible once they are labeled (Booher-Jennings, 2005). In some schools volunteers have been assigned to tutor the lowest performing students, freeing teacher time to work with the bubble students (Kleinman, West, & Sparkes, 1998). The practice of teaching to the bubble is more pronounced when schools focus on short term goals. When long term goals are emphasized there is increased hope for the poor performing students. However, when immediate returns are required remedial students that score well below passing require too much support and attention and are thus considered a liability (Reback, 2008).

Equity

The push for achievement tests refocuses attention on achievement and away from issues of equity (Gipps & Murphy, 1994). In the U.S. the well documented racial achievement gap has expanded with the implementation of increased accountability measures over the past two decades (Harris & Herrington, 2006). African American and Hispanic students continue to score substantially below their white peers across all subjects (Kornhaber, 2004a; McNeil, 2000; McNeil & Valenzuela, 2001; Wheelock, Bebell & Haney, 2000). English language learners (ELL) in the U.S. have also been disadvantaged by testing for accountability. In New York, ELL students went from having the highest minority graduation rate to the lowest in the time directly before and after NCLB (Monk,

Sipple, & Killeen, 2001). Finally, low performing students fare worse than high performing students, with the former exhibiting stagnant test scores and an increased likelihood of dropping out (Jacob, 2001; West & Pennell, 2000).

The differential outcomes of student groups can partially be tracked to the behavior of key actors in education. Teachers are more likely to hold lower expectations for low achieving students (Kornhaber, 2004a) and greater teacher turnover is found in low performing schools (Clotfelter et al., 2004; Figlio & Loeb, 2011; Waterreaus, 2003). The attrition in struggling schools is made worse because, on average, the more effective teachers leave (Feng, Figlio, & Sass, 2009). Student perception can influence student engagement and success in testing with poor motivation disproportionately found among low performing students (Paris & McEvoy, 2000). In a content analysis of 411 drawings from 4th, 8th, and 11th graders, Wheelock, Bebell, and Haney (2000), found that over 60% of students critiqued their state mandated test with urban students and minority students more likely to complain about test length, test difficulty, and hold overall feelings of hostility or anger toward the test. The actions of parents can enhance inequity when poor performing schools sink further as high achieving students are pulled out for higher performing schools (Hill, 2006). The problem of exit is captured in the dual role of students as the consumers of education and the producers of test scores. When families act as consumers their exit diminishes the chances of those left behind (Edwards & Whitty, 1992). Choice decisions in families are affected by their ability to interpret published school scores. Evidence from the UK suggests that approximately two-thirds of highly educated parents accurately interpret league tables while only 31% of less educated parents are able to do so (West & Pennell, 2000).

Validity

Opponents of testing for accountability believe accountability is an inappropriate aim of testing (Paris & McEvoy, 2000) with some suggesting testing has a sole purpose, to "serve as a tool to enhance all students' knowledge, skills, and understanding so that they can function at the highest possible level in the wider world" (Kornhaber, 2004a, p. 91). Standardized tests may not be able to capture the complex learning activities that measure critical understanding (Darling-Hammond, 1994; Froese-German, 2001). Standardized test scores have difficulty informing instruction due to longer feedback loops that make them inappropriate for formative assessment (Chapman & Snyder, 2000; Kornhaber, 2004a; Paris & McEvoy, 2000; Supovitz, 2009). Additionally, when information is provided concerns of "test score pollution" are present, with student motivation and family income acting as potential confounding variables (Froese-German, 2001; Hursh, 2007; Robertson, 1998; Wheelock et al., 2000).

Toward a Normative Testing Culture?

The global transformation toward testing for accountability should lead policymakers to question whether the potential benefits of implementing such a program outweigh the concerns, especially among the most marginalized student groups. Evidence suggests the magnitude of this effect, both positive and negative, increases in nations that apply Form Sanc/Rew policies, which use both market mechanisms and behavioral incentives. Unfortunately rich discussions of this nature are less likely among national decision makers in the future, as testing for accountability is increasingly legitimated as a neoliberal script, which lays out appropriate action for nation-states within a world culture that emphasizes faith in science, academic knowledge, and western style education. Testing as the way to

acquire valued knowledge is now taken for granted. As testing for accountability becomes a normative practice, engrained in the educational landscape of more and more countries, it has the potential to reconstruct education and how it is perceived by its actors and the general public.

Testing for accountability, as a neoliberal policy, reinforces the re-conceptualization of students, parents, and teachers, as products, consumers, and producers (Carl, 1994). With educator survival on the line, students are judged by their proclivity to pass a test. Remedial students, with long odds of passing the test, are considered a liability. The social construction of this student group as a 'liability' happens early in schooling and may follow the path of similarly constructed social categories that are now engrained in the rhetoric of education, 'dropouts' and 'at-risk' students (Baker, 2014; Fine, 1991; Swadener & Lubeck, 1995). In a schooled society, students that struggle in school will be identified as deviant and with education increasingly used as the only legitimate form of stratification 'liability' students will be marred by that status for the rest of their life (Baker, 2014).

As a centerpiece of testing for accountability systems, between school competition shifts "schools *modi operandi* from those based on moral purpose towards those that emphasize productivity and efficiency" (Sahlberg, 2010, p. 48). Once productivity and efficiency are established as the essential aim of schooling, testing for accountability policies may be criticized but will rarely be abolished (Tyack & Cuban, 1995). Competition between educators sharply contrasts with more cooperative models that are important for healthy school climates and student and teacher motivation. In systems where scores are aggregated at the class level teachers may concerned about peer judgment and being stigmatized as 'bad' teachers (Booher-Jennings, 2005). This tumultuous situation leads to educators blaming others, especially those in earlier grades, for not adequately preparing

students (Wiggins & Tymms, 2000). Internal feelings of anxiety and shame among teachers are exasperated by a belief that an emphasis on testing for accountability has a negative effect on public education, often stymieing their motivation (Certo, 2006; Jones & Egley, 2006; Smeed & Victory, 2010). Additionally, the focus on test scores suggests that everyone is welcome in teaching as long as they can produce the test gains needed in an accountability system (Hopmann, 2008). This assumption of importance threatens the professional position of teachers, delegitimizing it as a profession that requires long term pedagogical training. The importance of test score as information assumes that parents engage with the data and use it to drive their children's school placement. Parents that are uniformed, do not use evidence based decision making, or do not participate in school choice will be shunned by society (Ball, 1993). Evidence of subjective evaluations can already be found in the literature. For example, Woessman (2004) suggests that parents ability and willingness to make use of available information is a measure of "how strongly parents care for their children's progress" (p. 4).

The long term societal effect of treating education as a market, concerned primarily with private returns, is a reduction in public spending on social services, such as education and health (Hopmann, 2008). These reductions may be partially ameliorated by increased private support, accelerating the movement toward privatization. Evaluating quality in education is a challenging endeavor due to education's multifaceted short term and long term outcomes. Test results provide a simple, relatively easy to understand measure that can be used with investors who want to ensure their money is going towards a high quality product. Recent literature suggests this is the case, as the publication of results increase the amount of voluntary contributions a school receives (Figlio & Kenny, 2009). The publication of results can also shape society through residential segregation. Families use

school level test scores to judge the quality of the school system leading to the establishment of more 'highly desirable' neighborhoods and effecting real estate prices (Figlio & Lucas, 2004).

Testing for accountability is so engrained in many countries that it is partially self-perpetuating (Dorn, 2007). The use of assessment data reinforces the testing culture (Baker & Wiseman, 2005) and the public view of testing for accountability as synonymous with high expectations makes it challenging for policymakers to alter established practices, whether or not they want to, in fear of constituents labeling them soft on education (Paris & McEvoy, 2000). If the potential benefits and concerns of testing for accountability are not critically evaluated this global transformation will lead to a testing culture that is internalized as normative and adopted as individual values. This shift in constitutional mind-sets has the ability to affect the whole of society as "deeply engrained ways of understanding the relationship between the public and its institutions" (Hopmann, 2008, p. 425) are altered. Recognizing the cultural diffusion of testing for accountability and evaluating it before its cultural entrenchment is essential for the world to avoid the challenges faced by early adopters.

CHAPTER THREE: METHODOLOGY

This methodology chapter provides a step by step description of the planning and analytic decisions made in this study. The chapter starts with an introduction to the research questions that are driving the method selection. The primary data used in this study, taken from the 2009 Programme for International Student Assessment (PISA), is then described in detail including PISA sampling practices, and decisions regarding the elimination of cases and missing variable strategies. This is followed by a thorough breakdown of the included variables, including variable operationalization and source (if not from the 2009 PISA). The chapter continues with a description of the preliminary analysis used to investigate bivariate relationships. The intra-class correlation (ICC) and design effect statistics used to measure between cluster variance precedes the final section which describes the primary analysis in this study, hierarchical linear modeling (HLM). HLM, appropriate in this study given the presence of contextual measures in the research questions and the nested data structure, will be explained further in the last section of this chapter prior to specifying the models used in this study.

Research Questions

The global transformation of testing leads to an increase in education system scrutiny, assessing both quality and equity of educational outcomes. This study speaks to the potential benefits and concerns of testing for accountability by examining the relationship between national testing policy, contextual factors, and student outcomes. By examining national testing policy through a pooled sample of diverse countries, the results provided here can help identify common relationships with testing across heterogeneous environments. The examination of a phenomenon across different contexts can help elucidate its most common characteristics and strengthen causal inferences. Furthermore,

the inclusion of a rough coupling variable (described in greater detail later) may help direct policymakers to appropriate policy levers.

Assuming that the world culture has legitimated standardized testing and encouraged more countries to incorporate testing for accountability into their national policy, the research questions driving this study are threefold:

1. How is national testing policy related to the degree schools incorporate testing into their practices and policies?
2. Does the incorporation of testing into a school structure vary by the school's economic and/or academic composition?
3. How does the national testing policy and corresponding policy coupling influence student outcomes?

Data

Data from the 2009 PISA is the primary source of information for this study. The 2009 PISA was administered to 15 year olds in 61 countries and 4 sub-regions or participating economies by the Organization for Economic Co-operation and Development (OECD). The first round of PISA was administered in 2000. Since then PISA has been administered to a growing number of countries and economies every three years (see figure 1 for illustration of expansion). PISA targets 15 year olds to evaluate the ability of students near the end of compulsory school to apply their knowledge to meet real-world challenges (OECD, 2012). Since the assessment is not tightly linked to secondary school curriculum PISA is considered a soft test or low stakes test (Barry et al., 2010) and is more difficult to prepare for. Subjects covered on PISA include reading, mathematics, and science literacy, with one primary domain of interest covered in depth during each cycle. This study focuses on mathematics achievement because the stronger association between mathematics test

scores and accountability (Cronin et al., 2005) suggests that the analysis is more likely to find significant results.

Assessment scores are standardized around an OECD country mean score of 500 with a standard deviation of 100. Given the limited time for the assessments students are unable to answer all assessment questions. Therefore, student level assessment scores are provided through an imputation methodology known as plausible values. Each student is provided five plausible scores based on their responses to questions of similar difficulty and discrimination using Item Response Theory (OECD, 2012).

Two stage stratified sampling is conducted to create a representative sample of each country's 15 year old population. A minimum of 150 schools are randomly chosen from all public and privately controlled schools. Thirty five students from all programs (academic, vocational, etc.) within the chosen school are then randomly sampled. If less than 35 students are present, all students are selected. The minimum total sample of students per country is 4,500 with countries allowed to exclude up to 5% of their 15 year old population due to inability to take the test (due to functional, intellectual, or language difficulties) or administer the test (often due to very remote rural school location) (Hopstock & Pelczar, 2011). Approximately 470,000 students participated in the 2009 PISA representing 26 million 15 year olds across all participating countries and economies (OECD, 2012). In addition to taking the assessment students complete a student questionnaire focused on their background, learning habits, and attitudes toward learning. Additional contextual information is provided through school questionnaires completed by the principal which ask questions about student demographics and school environment.

The sample for this study pools data from all 2009 PISA participating countries and economies. Cases missing dependent variables (i.e. school practice and student

achievement) are omitted from this analysis. Additionally, nested units (i.e. schools) with two or fewer cases (i.e. students) are omitted from analysis to prevent the overestimation of explained variance and the exaggeration of model fit (Subedi, 2005). Following the omission of these cases the remaining missing data is addressed through list-wise deletion. With exclusion criteria and missing data procedures taken into account, the effective sample size from this study is 446,330.

Variables

Variables for this study are included at three levels of analysis: nation, school, and student. The following section describes variables at each level including how they are operationalized and coded. For a complete list of all variables see Table 3. While the majority of variables are drawn from the 2009 PISA student and school questionnaire, those taken from other sources (including the entirety of national level variables) are clearly indicated.

Table 3: All Included Variables by Name and ID

National Level Variables				
Variable Name	**Variable Type**	**Range**	**Variable ID**	**Centering**
National Testing Policy	Categorical	0-3	ntp	
National Testing Policy: No NTP	Dummy Coded	0-1	no_ntp	
National Testing Policy: Summative	Dummy Coded	0-1	summative	
National Testing Policy: Evaluative	Dummy Coded	0-1	evaluative	
National Testing Policy: Form Sanc/Rew	Dummy Coded	0-1	form_sanc_rew	
Economic Development	Continuous	-22872-45623	n_develop_c	grand mean
Education Expenditure	Continuous	2.0-8.1	n_ed_expend	
Market Competition	Continuous	0.08-96.07	n_comp	
School Level Variables				
Variable Name	**Variable Type**	**Range**	**Variable ID**	**Centering**
AC: School Monitoring	Dichotomous	0-1	sc_monitor	
AC: School Comparison	Dichotomous	0-1	sc_comp	
AC: Parent-School Comparison	Dichotomous	0-1	p_sc_comp	
AC: Publicly Posted	Dichotomous	0-1	posted	
AC: Principal Evaluation	Dichotomous	0-1	pr_eval	
AC: Teacher Evaluation	Dichotomous	0-1	t_eval	
School Compliance	Continuous	0-6	compliance	
Admision Decision	Dichotomous	0-1	admin	
Transfer Decision	Dichotomous	0-1	transfer	
Standardized Tests (>2/year)	Dichotomous	0-1	s_tests	
Extra-curricular Activities	Continuous	0-13	ex_act	
School Mean Math Score	Continuous	-338.42-245.13	sc_math_c	grand mean
School SES	Continuous	-4.20-1.75	sc_ses	
School Type: Private	Dichotomous	0-1	sc_private	
School Location: Small Town	Dummy Coded	0-1	sc_loc_smalltown	
School Location: Town	Dummy Coded	0-1	sc_loc_town	
School Location: City	Dummy Coded	0-1	sc_loc_city	
School Location: Large City	Dummy Coded	0-1	sc_loc_largecity	
Student Level Variables				
Variable Name	**Variable Type**	**Range**	**Variable ID**	**Centering**
Student Math Achievement	Continuous	3.67-890.10	PV1MATH-PV5MATH	
Gender	Dichotomous	0-1	female	
SES	Continuous	-6.62-3.44	ses	
Home Language	Dichotomous	0-1	lang_other	
Immigrant Status	Dichotomous	0-1	immig	
Grade Repetition	Dichotomous	0-1	repetition	
Time in Math	Continuous	-225.53-2174.47	time_math_c	grand mean
CL: School Prepares me for Life	Dichotomous	0-1	cl_prep_life	
CL: School is not a Waste of Time	Dichotomous	0-1	cl_not_waste	
CL: Teacher is Interested in my Well-Being	Dichotomous	0-1	cl_t_interested	
CL: Teacher Treats me Fairly	Dichotomous	0-1	cl_t_fairly	

National Level Variables

National testing policy (NTP) is identified through a careful examination of national and international policy documents. Policies were categorized as of the 2008-2009 school year and dummy coded using the four categories identified in chapter two: Formative,

Summative, Evaluative, and Form Sanc/Rew. In countries with multiple examinations the examination with higher levels of accountability (i.e. toward Form Sanc/Rew) are included as the country example. The steps taken in the categorization process are laid out in Figure 4. When possible multiple sources were used to triangulate the data, increasing categorization accuracy.

Figure 4: Country Categorization Scheme

Categorization Step	Topic	Explanation	Coding	Example
1	Presence of National Testing Policy	Is a national testing policy present that indicates the use of a national or regional examination?	If yes = proceed to step 2 If no = No National Testing Policy	
2	Aggregation and Dissemination	What level (class, school, regional, national) are the results aggregated and shared with the public?	If not shared with the public = Formative Testing Policy If class or school aggregation = proceed to step 3 If regional or national = Summative Testing Policy	Summative = Traditional high stakes student exams for advancement (Germany)
3	Explicit Consequences	If aggregated at the school/class level are sanctions or rewards attached to student test scores.	If no sanctions/ rewards = Evaluative Testing Policy If sanctions/rewards = Form Sanc/Rew Testing Policy	Evaluative = League tables (United Kingdom) Form Sanc/Rew = No Child Left Behind (United States)

Following initial categorization national experts from within the researcher's network were consulted to clarify ambiguous country categories. An updated country

categorization was then sent out to the OECD Group of National Experts on Evaluation and Assessment for review. Their feedback provided additional support for the categorization through the identification of individual country nuances. Countries with large regional variance in their NTP are coded into multiple categorizes when the variance divides the country by NTP and data available in the 2009 PISA can be divided by region (i.e. Belgium, Canada, and the United Kingdom). Table 4 outlines the complete country categorization grid for 2009 PISA participants. Note that the Formative and Summative categories have been collapsed due to the limited amount of countries that keep test results entirely internal and the challenges in distinguishing between the categories within a single policy. When available, in countries where NTP varies by region, regional information is use to disaggregate national scores. See Appendix A for sources used in NTP categorization and Appendix B for representative references of each country in the testing for accountability categories.

Table 4: National Testing Policy Categories based on ISCED 1 and ISCED 2 Education Policies[1]

No National Testing Policy Identified at the ISCED 1 or ISCED 2 level	Summative Testing Policy	Evaluative Testing Policy	Formal Sanction and/or Reward Testing Policy
Czech Republic[7]	Albania[37, 57]	Australia[7, 11]	Chile[5, 6, 7, 58]
Dubai-UAE[51]	Argentina[5]	Azerbaijan[49, 50]	Hong Kong-China[3, 4]
Greece[2, 7]	Austria[1, 2, 7, 10]	Belgium (Flemish)[1, 7, 8]	Hungary[1, 2, 7]
Liechtenstein[1, 23]	Belgium (French and German)[1, 7, 8]	Brazil[3, 4, 5, 12]	Latvia[1, 22, 34]
Panama[5]	Bulgaria[1, 7, 13, 14]	Canada (Ontario)[3, 4]	Macao-China[3, 4]
Switzerland[7, 45, 46]	Canada (Other)[3, 4]	Columbia[5, 15, 32]	Mexico[5, 7, 8]
	Chinese Taipei[40]	Denmark[1, 7, 8]	Portugal[1, 2, 7]
	Croatia[35]	Estonia[1, 2, 7]	Shanghai-China[3, 4]
	Finland[1, 7, 16]	Iceland[1, 7, 17, 18]	United States[7, 55]
	France[1, 3, 4, 7, 11]	Korea, Republic of[3, 4, 7]	
	Germany[1, 3, 4, 7]	Lithuania[1, 2, 24]	
	Indonesia[7]	Netherlands[1, 3, 4, 7]	
	Ireland[1, 2, 7]	Norway[1, 7, 8]	
	Israel[7, 19]	Poland[1, 2, 7, 59]	
	Italy[1, 2, 7, 20, 33]	Qatar[38, 39]	
	Japan[7, 21]	Romania[1, 2, 25]	
	Jordan[43, 44]	Singapore[3, 4]	
	Kazakhstan[48]	Slovenia[1, 2]	
	Kyrgyz Republic[36]	Sweden[1, 2, 7]	
	Luxembourg[1, 7, 8]	Trinidad and Tobago[52]	
	Montenegro[37]	Turkey[7, 31, 56]	
	New Zealand[7, 8, 9]	United Kingdom [1, 2]	
	Peru[5]		
	Russian Federation[3, 4, 7, 26, 27, 28]		
	Serbia[37, 47]		
	Slovak Republic[1, 7]		
	Spain[1, 2, 7, 30]		
	Thailand[41, 42]		
	Tunisia[53, 54]		
	Uruguay[5]		

Economic development is a national level control variable, operationalized as the GDP per capita (PPP). Data on economic development comes from the World Bank Databank. To smooth out economic shock from the onset of the global recession in 2008, a five year (2005-2009) national average is used.

Education expenditure is a continuous national level control variable that is used as a proxy for national investment in education. Data on education expenditure is taken from the

[1] Superscript number corresponds with source used for categorization (see Appendix A)

World Bank Databank. To ensure that education expenditure accurately represents national investment in education, given the global recession, a five year average of public expenditure in education as a percent of GDP is used.

Market competition is a continuous national level control variable that is operationalized as the percentage of total enrollment that attends a private school. Market competition, therefore, acts as a proxy for competition between schools. Data for the market competition variable is taken from the World Bank Databank with the 2009 national percentage of private school enrollment at the elementary level averaged with the 2009 percentage at the lower secondary level.

School Level Variables

Six school accountability variables are taken from the 2009 PISA school questionnaire to identify within school practices. These variables are then combined to form a school compliance variable. All school accountability variables are dichotomous (1 = yes, 0 = no). The first two accountability variables are taken from question number 16 of the school questionnaire. The first variable, school monitoring, is taken from the question: "Are assessments of students used to monitor the schools progress from year to year?" The second variable, school comparison, asks the question: "Are assessments of students used to compare the school with other schools?" From question number 21 of the questionnaire, parent-school comparison addresses the question: "Does your school provide aggregated and comparable school results to parents?" The last three school accountability variables are taken from question number 22 and include: publicly posted ("Is achievement data posted publicly?"), principal evaluation ("Is achievement data used to evaluate the principal's performance?"), and teacher evaluation ("Is achievement data used to evaluate the teacher's performance?").

Local level practice may vary from national level policy. This is especially true in education where local policy enactors, often termed street level bureaucrats, apply the policy differently depending on their beliefs, experiences, and abilities (Datnow & Park, 2009; Shulman, 1983). To capture within country heterogeneous implementation of NTP a school compliance variable is created from the six previously identified school accountability variables. As expected school practices differ by specific NTP, a school that is fully compliant in one national context may be non-compliant in another. For example, a school that publicly posts its test scores would be compliant in an Evaluative or Form Sanc/Rew testing policy but non-compliant in a Summative testing policy. To create a variable that captures the different expected practices given NTP a school compliance variable matrix was used (see Table 5). Using the matrix one point is given to schools for each category the school practice is "present in compliant school" given its NTP, thus the school compliance variable has a range of zero to six. For example, if a school is the U.S. (Form Sanc/Rew testing policy) answered yes (present) to all school accountability variables but principal evaluation they would have a school compliance score of five out of six (83.3%), indicating very strong but not complete compliance.

Table 5: School Compliance Matrix

	School Monitoring	School Comparison	Parent-School Comparison	Publicly Posted	Principal Evaluation	Teacher Evaluation
Summative Testing Policy	**O**	**O**	**O**	**O**	**O**	**O**
Evaluative Testing Policy	**X**	**X**	**X**	**X**	**O**	**O**
Form Sanc/Rew Testing Policy	**X**	**X**	**X**	**X**	**X**	**X**

O = Absent in compliant school
X = Present in compliant school

Additional school level practices are included to address policies that may be used by schools to game the system, artificially inflating school aggregate achievement scores without increasing the quality of instruction or education. To identify whether a school may be shaping their testing pool be selecting in higher achieving students while selecting out low achieving students, schools with selective admission and transfer policies where identified from the school questionnaire. The former identifies schools where student achievement, including test scores, are included in the admission criteria (1= Sometimes or Always; 0 = Never). The later identifies whether a student is likely to be transferred out due to low academic achievement (1 = Likely to Very Likely; 0 = Not Likely).

When under pressure schools may be more likely to teach to the test. To capture whether schools practice taking standardized tests a dichotomous standardized test variable is created from question 15 of the school questionnaire. Schools that complete two or more standardized tests a year are considered high tested schools and coded as 1. Schools that complete fewer than two standardized tests annually are considered low tested schools and coded as 0. Additionally, to identify whether schools in testing for accountability systems offer less extra-curricular activities, thus availing more resources for tested subjects, a composite extra-curricular activity variable is created by adding the number of positive responses to question 13 of the school questionnaire. The composite ranges from none of the extra-curricular activities are available at the school (0) to all of the extra-curricular activities are offered at the school (13).

To capture the confounding potential of school characteristics, four school level control variables are used. First, school math achievement is a continuous variable calculated by aggregating within school student math achievement. Second, school socio-economic status (SES) is a continuous variable calculated by aggregating within school

student index of economic, social and cultural status (ESCS, explained in the student level variable section). The final two school level control variables are taken from the 2009 PISA school questionnaire. School type is a dichotomous variable derived from the question: "Is your school public or private"? (1 = private, 0 = public). School location is a dummy coded variable that addresses the question: "Which of the following best describes the school community?" Dummy coded location variables include: small town or village (population less than 15,000), town (population 15,000 to 100,000), city (population 100,000 to 1,000,000), and large city (population greater than 1,000,000).

Student Level Variables

Student mathematics achievement scores (hereafter student achievement) is derived from the five plausible values provided in the 2009 PISA. To combine plausible values student achievement mean is calculated from the five scores and the standard errors are adjusted according to the rules of Rubin (1987) to take into account their imputed nature (OECD, 2012).

Additional student level outcome variables include: grade repetition, time spent in mathematics, and four variables that capture student perception of the school climate. Grade repetition is a dichotomous variable taken from question number 7 of the student questionnaire: "Have you ever repeated a grade in ISCED 1 or ISCED 2?" and coded 1 for yes and 0 for no. Time spent in mathematics is a continuous variable taken from the product of the student's report of how many minutes are spent in every math class (student questionnaire #28) and how many math classes do you have per week (student questionnaire #29). Student perception of school climate variables are dichotomous variables that have been reverse scored when necessary to measure whether a student agrees or strongly agrees to the following statements: School has prepared me for adult life when I leave school

(student questionnaire #33a); school has not been a waste of time (#33b); most of my teachers are interested in my well-being (#34b); and most of my teachers treat me fairly (#34e).

Student level control variables include gender, SES, home language, and immigrant status. Gender is a dichotomous variable coded 1 for female and 0 for male. SES is taken from the PISA created student level index of economic, social and cultural status (ESCS). ESCS is a composite variable created through principle component analysis. Included in the composite are variables that identify home possessions, the highest parental occupation, and the highest parental education in years of education. It is a standardized variable with an OECD mean of zero and standard deviation of one. Median factor loadings for the observed variables in OECD countries ranged from 0.70 (home possessions) to 0.80 (parental education). Median scale reliability for OECD countries was 0.65 (OECD, 2012).

Home language is a dichotomous control variable taken from question number 19 of the student questionnaire: "What language do you speak at home most of the time?" The variable is coded 1 if the student does not speak the test language as their primary home language and 0 if the test language is their primary home language. Immigrant status is taken from the PISA created immigrant status variable and coded 1 if the student is a first or second generation immigrant (i.e. they were not born in the country they were tested) and 0 if they are native to the test country.

Descriptive Statistics

Table 6 provides descriptive statistics for all variables by NTP as well as the overall sample, allowing for the identification of compositional differences between different testing policies. Standard deviations are provided in parentheses. The table suggests Evaluative and Form Sanc/Rew systems are generally similar across national, school, and student levels.

Systems that do not employ a national or regional test tend to be found in higher income countries with more private enrollment.

Table 6: National, School, and Student Level Descriptive Statistics

National Level Variables

	Overall	**No NTP**	**Summative**	**Evaluative**	**Form Sanc/Rew**
Economic Development	$24,762.85 (14,697.68)	$35,615.48 (13,890.50)	$22,605.05 (13,244.57)	$27,358.44 (16,118.59)	$20,487.73 (12,415.56)
Education Expenditure	4.70% (1.10)	4.57% (0.67)	4.55% (0.98)	4.95% (1.35)	4.65% (0.88)
Market Competition	15.95% (18.75)	22.16% (26.21)	12.64% (13.09)	17.26% (17.23)	19.20% (26.27)
N	66	6	29	22	9

Note: Standard deviations in parentheses.

School Level Variables

	Overall	**No NTP**	**Summative**	**Evaluative**	**Form Sanc/Rew**
School Achievement	467.72 (78.77)	476.70 (82.50)	465.66 (78.46)	467.87 (79.80)	468.72 (75.63)
School SES	-0.27 (0.78)	0.03 (0.56)	-0.23 (0.72)	-0.15 (1.11)	-0.73 (0.84)
School Type – Private	18.48%	22.12%	16.40%	17.08%	24.76%
Location: Small Town	30.14%	41.52%	31.54%	28.59%	24.03%
Location: Town	30.41%	28.31%	34.72%	29.30%	22.04%
Location: City	24.94%	17.35%	23.38%	25.61%	31.32%
Location: Large City	14.51%	12.82%	10.34%	16.50%	22.61%
AC: School Monitoring	82.22%	67.81%	78.84%	87.06%	89.06%
AC: School Comparison	50.19%	41.38%	43.56%	57.11%	59.32%
AC: Parent-School Comparison	25.03%	24.46%	21.62%	28.53%	28.01%
AC: Publicly Posted	34.84%	23.54%	29.95%	44.75%	35.27%
AC: Principal Evaluation	39.65%	31.74%	34.28%	50.34%	38.42%
AC: Teacher Evaluation	54.71%	49.72%	47.14%	61.55%	64.69%
School Compliance	3.26 (1.57)	NA	3.45 (1.80)	3.06 (1.12)	3.15 (1.55)
Admission Decision	57.37%	69.46%	55.84%	51.25%	66.66%
Transfer Decision	35.30%	36.52%	39.11%	24.32%	44.61%
Standardized Tests (>2/year)	30.24%	30.95%	26.86%	32.21%	35.25%
Extracurricular Activities	7.60 (2.90)	8.00 (2.54)	7.43 (2.84)	7.72 (3.04)	7.64 (2.88)
N	15,785	1,402	6,919	4,845	2,619

Note: Standard deviations in parentheses.

Student Level Variables

	Overall	**No NTP**	**Summative**	**Evaluative**	**Form Sanc/Rew**
SES	-0.27 (1.13)	0.03 (0.93)	-0.23 (1.09)	-0.15 (1.11)	-0.73 (1.23)
Female	50.61%	49.41%	50.44%	50.85%	51.12%
Home Language – Other	11.47%	17.24%	14.43%	9.94%	4.03%
Immigrant Status	10.54%	25.77%	8.16%	9.99%	10.84%
Time Spent in Math	225.53 (99.43)	232.22 (91.44)	218.75 (98.16)	218.48 (99.16)	252.45 (101.69)
Grade Repetition	10.96%	12.02%	9.99%	10.72%	13.32%
CL: Prepares for Life	73.61%	68.51%	73.85%	76.24%	70.67%
CL: Not a Waste	92.33%	90.54%	92.03%	92.02%	94.46%
CL: Teacher is Interested	71.74%	73.17%	69.74%	71.78%	76.19%
CL: Treats me Fairly	80.19%	78.27%	80.66%	81.40%	77.73%
N	446,330	34,801	199,998	135,157	76,374

Note: Standard deviations in parentheses.

Preliminary Analysis

Prior to conducting the primary multi-level analysis, preliminary analyses are conducted to explore the bivariate relationship between key variables. To test for significant differences by NTP chi-square tests and one way analysis of variance (ANOVAs) are conducted with appropriate post-hoc tests to clarify the bivariate results. One way ANOVAs investigate differences in the following continuous variables by NTP: school SES, school achievement, school compliance, economic development, education expenditure, and market competition. Bartlett's test for equality of variances tests for equal variances across groups and the Sidak multiple comparison test are used post hoc to identify the specific groups that differ (Kremelberg, 2010).

Pearson chi-square goodness of fit tests are used to compare the relationship between two dichotomous or categorical variables to the hypothesized equal distribution in each cell

(Kremelberg, 2010). To investigate whether the observed frequencies vary from the hypothesized cell size separate chi-square tests are ran between NTP and all school accountability variables.

To test whether some national level characteristics are likely to be related to NTP a multinomial regression is conducted with NTP as the dependent variable and all national level variables included as predictor variables. Significant results in the predictor variables would suggest large between country variance and justify the inclusion of such variables as control variables in the primary multi-level analysis.

Calculating Between Group Variance

Adequate between group variance is a necessary condition to ensure the utility of multi-level models. Even when data is nested if no variance is attributed to the higher level units multi-level modeling is not necessary and acts in equivalence to multiple regression (Peugh, 2010). Two statistics are used to measure the amount of variance present between higher level units: the intraclass correlation (ICC) and the design effect.

The ICC measures the proportion of variation in the outcome variable that can be attributed to differences between higher level units. A large ICC is indicative of large between group variance, small within group variance, or a combination of the two. To calculate the ICC an unconditional multi-level model or empty model is ran (O'Connell, & Reed, 2012; Peugh, 2010). An unconditional model contains no predictor variables and is akin to a one way ANOVA (Subedi, 2005).

Equation 1 depicts the unconditional multilevel model. Variable *Y* represents the dependent variable (i.e. student achievement) for individual *i* in group *j*. The dependent variable is modeled as a function of the intercept, β, plus an individual specific deviation, *e*, and a group specific deviation, *u*. The unconditional model does not explain what causes the

variance in *Y*, instead it partitions variance into within group and between group components (Maas & Hox, 2005).

Unconditional Multilevel Model: $Y_{ij} = \beta_{0j} + e_{ij} + u_{0j}$ (Equation 1)

Equation 2 uses the individual and group specific deviation present in equation 1 to calculate the ICC. In the equation the within group variance is denoted as σ^2_e while the between group variance is denoted as σ^2_u. The ICC then measures the proportion of variance that occurs between groups (Peugh, 2010). Common ICC values for social science research range from .05 to .20 (Peugh, 2010). Typical data using student achievement as the dependent variable has seen ICCs in the .25 to .30 range (Lee, 2000). Trivial ICC, indicating that multi-level modeling is unnecessary, is suggested at less than .10 (Lee, 2000). Additionally, the ICC value may be less informative in model interpretation when a binary variable is used as the dependent variable as lower level variance in non-linear models is heteroscedastic (Subedi, 2005).

$ICC = \sigma^2_{u0}/(\sigma^2_{u0} + \sigma^2_e)$ (Equation 2)

The design effect "represents how much the standard errors from a clustered design are underestimated relative to a simple random sample" (O'Connell & Reed, 2012, p. 18) and is calculated using the ICC. A simple random sampling design has a design effect of 1.0 while a design effect over 2.0 indicates a need for multi-level modeling (Muthen, 1994). Equation 3 illustrates the design effect with n_c denoting the average cluster size.

$$\text{Design Effect} = 1 + (n_c - 1)\text{ICC} \qquad \text{(Equation 3)}$$

Primary Analysis

Method

The presence of non-trivial between group variance reinforces the need for a multi-level model. In education research multi-level models are commonly known as hierarchical linear models (HLMs). HLM acknowledge the nested, or hierarchical, nature of data (Raudenbush & Bryk 2002), making it the appropriate method for this study where students are nested in schools which are nested in countries. Traditional approaches that fail to take into account the hierarchical data structure violate the independence of observations assumption (Bryk & Raudenbush, 1992; Osborne, 2000; Peugh, 2010). As individuals in groups tend to be more similar than individuals randomly sampled from the entire population, a relative measure of within group interdependence is present. This may be due to within group interaction, contextual factors that lead to shared group experiences, or non-randomly distributed family background variables (Krull & MacKinnon, 2001). This violation suggests that multiple regression estimations generate standard errors that are too small, increasing Type I error. Additionally, between group variance violates the constant variance assumption of multiple regression (Bryk & Raudenbush, 1992; Hofmann & Gavin, 1998).

In the past research has often attempted to overlook or discard the hierarchical data structure through the aggregation or disaggregation of variables, both of which change the nature of the research question (Osborne, 2000). The aggregation of lower level data to the higher level in order to make inferences about higher level outcomes is known as atomistic fallacy (O'Connell & Reed, 2012) and results in a substantial decrease in the variance on the outcome variables as well as the power of statistical tests (Krull & MacKinnon, 2001). The disaggregation of higher level data to the lower level in order to make inferences about lower level outcomes is known as ecological fallacy (O'Connell & Reed, 2012) and results in the violation of independent observations (Krull & MacKinnon, 2001).

HLM overcomes these prior problems by correcting for the bias resulting from the hierarchical data structure and providing correct standard errors that can be used in significance tests or to calculate confidence intervals (CIs) (Guo & Zhao, 2000). This is accomplished through the decomposition of variance into within group and between group components and HLMs ability to incorporate cross level interactions into the model (Bryk & Raudenbush, 1992; Krull & MacKinnon, 2001). The advantages of HLM, combined with the 2009 PISA hierarchical data structure and the research questions previously outlined, make it the appropriate method for this study.

Steps in Analysis

Step 1: Model Estimation

To conduct HLM analysis this study roughly follows the steps outlined by Peugh (2010). Once research questions are clarified and the use of HLM is justified through the presence of non-trivial between group variance, the appropriate parameter estimate is chosen. For HLM these decisions are generally limited to the choice of two maximum likelihood (ML) algorithms: "Maximum likelihood techniques provide estimates for the

values of the population parameters that maximize the probability of obtaining the observed data" (McCoach & Black, 2012, p. 24). Each ML algorithm provides possible values provided the given model parameters and when the maximum likelihood has been reached the model is said to have converged (Dedrick et al., 2009). ML is appropriate provided this study's large sample (Maas & Hox, 2005) which overcomes the 50/20 rule, 50 higher level groups with an average of 20 lower level units per group, suggested by Hox (1998) for HLMs that include cross level interaction.

The two common ML algorithms used in education research are full information maximum likelihood (FIML) and restricted maximum likelihood (REML). FIML include both regression coefficients and the variance components into the likelihood function while REML includes only variance components (Hox, 1998; Peugh, 2010). FIML is used when post hoc likelihood ratio tests are ran since both coefficients and components are necessary for the tests. In addition to this benefit, FIML is chosen in this study because it tends to be easier to compute and differences between the two algorithms are negligible with large sample sizes (Hox, 1998).

Step 2: Centering

Centering is an important to consider when conducting HLM because lower level coefficients are explained through higher level models (Dedrick et al., 2009). Additionally centering can help reduce multicolinearity between the random intercepts and random slopes included in the model (Kreft, De Leeuw, & Aiken, 1995; Krull & MacKinnon, 2001). Centering is necessary when the value of zero has no substantive meaning, making it difficult to interpret. For example, although PISA reading achievement scores are reported on a scale of 0 to 1000, in practice a score of zero is meaningless as only 1% of students in OECD countries score below the 1b proficiency threshold of 262 (OECD, 2012). Centering

then involves "rescaling a predictor variable so that a value of zero can be interpreted meaningfully" (Peugh, 2010, p. 91). Centering is not necessary for dichotomous variables (Peugh, 2010) but should be used for continuous lower level predictor variables (O'Connell & Reed, 2012).

Two approaches to centering are common: grand mean centering and group mean centering. In grand mean centering individual scores are subtracted from the overall sample mean (i.e. Achievement – Sample Mean Achievement). It is the only centering that can be used on higher level variables and reduces multicollinearity issues when cross level interaction terms are included (Raudenbush & Byrk, 2002). When grand mean centering is employed the lower level intercept can be interpreted as the expected value of an individual with the sample mean across all groups. In group mean centering individual scores are subtracted from the mean score of the individuals within the same higher level group (i.e. Achievement –Group Mean Achievement$_j$). When applied the lower level intercept is interpreted as the expected value of an individual at the mean of a given group (Hofmann & Gavin, 1998).

In this study grand mean centering is applied to continuous lower level predictor variables because it adjusts group variance difference by the lower level predictors, specifically the "level 2 regression coefficients represent the group level relationship between the level 2 predictors and the outcome variable less the influence of the level 1 predictor(s)" (Hofmann & Gavin, 1998, p. 628). This approach is appropriate because this study is most interested in the higher level or contextual effects on lower level outcomes where the lower level variables act primarily as control variables (McCoach & Black, 2012).

Step 3: Model Creation

From the unconditional model used to capture the ICC and design effects, lower level predictor and control variables are used to create the individual, base level, or within group model (Osborne, 2000). With only a single level considered, the results of the level 1 model are analogous to multiple regression. In equation 4 outcome variable (*Y*) for individual *i* in group *j* is predicted through the addition of an intercept (β_0), a predictor variable (β_1), such as SES, and a random error (*e*).

Level 1 Conditional Model: $Y_{ij} = \beta_{0j} + \beta_{1j}(SES) + e_{ij}$ (Equation 4)

The addition of level two variables helps capture contextual effects: "the main effects of higher level variables on outcomes defined at a lower level after controlling for relevant individual level cofounders" (Diez Roux, 2002, p. 589). Five assumptions are present in HLM including: error terms of each level should be normally distributed with a mean of zero, the relationship between the predictor and the outcome variable is linear, there is homogeneity of error variances at each level, at each level error terms are independent of predictors, and error terms between levels are independent (Subedi, 2005). At the second level, level one intercepts and coefficients are modeled as outcomes of level two predictor variables. A random intercept model or intercept as outcome model is present when only the level one intercept is modeled with a level two error term. Equation 5 illustrates how a level one intercept (β_0) can be predicted by a level two intercept (γ_{00}), predictor variable (γ_{01}), such as class size, and error term (u_0). In a random intercept model the level one intercept varies by group *j*, adjusting the intercept for each individual *i*, relative to their group membership and illustrating between group differences.

Level 2 Random Intercept: $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (Class Size) + u_{0j} (Equation 5)

To identify within group differences the regression coefficient of a fixed level one predictor variable is relaxed. The level one coefficient (β_1) is then modeled as an outcome of a level two intercept (γ_{10}), predictor variable (γ_{11}), such as school type, and error term (u_1), see equation 6. Random slope models create cross level interactions that modify the effects of a lower level variable by the characteristics of the higher level variable in which it is grouped in (Diez Roux, 2002). Level two error terms are important in both the random intercept and random slope model as they allow for the modeling of between group differences.

Level 2 Random Slope: $\beta_{1j} = \gamma_{10} + \gamma_{11}$(School Type) + u_{1j} (Equation 6)

By substituting equation 5 and equation 6 in for the appropriate intercept and coefficient in equation 4 we get the complete model displayed in equation 7. By rearranging the variables we can see that the segment [$\gamma_{00} + \gamma_{01}$ (Class Size) + γ_{10} (SES) + γ_{11}(School Type * SES)] contains all the fixed intercepts and coefficients; thus it is the fixed part of the model. The segment [$u_{0j} + u_{1j}$ (SES) + e_{ij}] contains all the random error terms and is thus the random part of the model.

Complete 2 Level Random Slope Model:

$$Y_{ij} = \gamma_{00} + \gamma_{01} \text{ (Class Size)} + \gamma_{10} \text{ (SES)} + \gamma_{11} \text{(School Type * SES)} + u_{0j} + u_{1j} \text{ (SES)} + e_{ij}$$

(Equation 7)

When the level one outcome variable is binary the normality of errors assumption is violated (Dedrick et al., 2009; Raudenbush & Byrk, 2002), therefore a hierarchical generalized linear model (HGLM), which relaxes the assumption, is used. Conceptually HGLM is similar to HLM with the logit function of the outcome variable as the primary difference (Guo & Zhao, 2000). The logit function provides the log odds of a binary outcome. The odds of an outcome are defined as the ratio of the probability of the outcome occurring (p_{ij}) relative to probability of the outcome not occurring ($1\text{-}p_{ij}$). The odds ratio (OR) can be interpreted as how many times more or less likely is the outcome (dependent variable) to occur in a given group relative to the same outcome occurring in the reference group (Subedi, 2005). The logit function transforms the odds into log odds by taking the natural logarithm of the odds (see equation 8). HGLM differs in estimation method from HLM, applying either marginal quasi-likelihood (MQL) or penalized quasi-likelihood (PQL). The change in estimation method can lead to problems with model convergence (Guo & Zhao, 2000).

Level 1 Model with Logit Function: $\log [p_{ij}/(1 - p_{ij})] = \beta_{0j} + \beta_{1j}(SES) + e_{ij}$ (Equation 8)

The inclusion of a third level is uncommon in education research (Subedi, 2005). For a three level model, level one coefficients are outcomes in the level two model and the level two coefficients are outcomes in the level three model. Equation 9 predicts the level two intercept (γ_{00}) from a level three intercept (δ_{000}), predictor variable (δ_{001}), such as national policy, and error term (ν_{00k}). Equation 10 substitutes equation 9 into equation 7, predicting the outcome variable for the *i* individual in the *j* group in the *k* context. Difficulty

in interpreting three level HLM amplifies the importance of balancing complexity with parsimony (McCoach & Black, 2012).

Level 3 Random Intercept: $\gamma_{00k} = \delta_{000} + \delta_{001}$ (National Policy) + ν_{00k} (Equation 9)

Complete Level 3 Random Slope Model:

$Y_{ijk} = \delta_{000} + \delta_{001}$ (National Policy) + γ_{01k} (Class Size) + γ_{10k} (SES) + γ_{11k}(School Type * SES) + $\nu_{00k} + u_{0jk} + u_{1jk}$ (SES) + e_{ijk} (Equation 10)

Step 4: Reporting Effect Size

The effect size reported is dependent on whether variance captured by the overall model is of interest (global effect size) or variance contributed to single variables within the model is of interest (local effect size). Although the r-squared statistic usually used to compute the global effect size in multiple regression analysis is unavailable in HLM, a pseudo r-squared has been proposed to identify the overall variance in the outcome variable captured by the complete model (see Singer & Willett, 2003; Snijders & Bosker 1994). Local effect size is calculated through the proportional reduction in variance (PRV) statistic and confidence intervals (CI). The PRV statistic was suggested by Raudenbush & Byrk (2002) and calculates the relative reduction in variance of a model after a single variable is added. Equation 11 lays out this computation where "var" represents the level one variance, level two slope variance, or level two intercept variance. The "prior" subscript indicates the level of variance estimated from the model prior to adding the variable and "predictor" indicates the level of variance estimated from the model after adding the variable.

$$PRV = (var_{prior} - var_{predictor})/var_{prior} \quad \text{(Equation 11)}$$

Due to the large sample size of this study CIs provide greater insights on the magnitude of results. Instead of simply discussing whether results are significant (as significant results are more likely with large sample sizes), CIs address whether the significant results are substantial. CIs can also be used as a measure of precision, a small CI indicates greater precision due to the small standard error (McCoach & Black, 2012). For a significance level of .05, a 95% CI is calculated. The endpoints of a 95% CI are the product of (+/-) 1.96 times the standard error plus the regression coefficient estimate. A 95% confidence interval indicates that 95% of the time we can be certain that true effect lies between the two endpoints (i.e. the range).

Step 5: Model Evaluation

To evaluate and contrast models a chi-square likelihood ratio test and the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) is used. The likelihood test can be used to test for statistical differences between two nested models. Deviance values are computing using -2 times the log likelihood of the model and the difference in deviance values is used as a chi-square distribution to calculate the test statistic (Peugh, 2010). In large samples the difference in deviance values approximates a chi-square distribution with the degrees of freedom equal to the difference in the number of parameters between models (McCoach & Black, 2012). Additionally, in such samples the Wald z test and the likelihood ratio test is equivalent. In likelihood ratio tests the more parsimonious model is preferred unless the change in deviance indicates it is a significantly worse fit.

The AIC and BIC evaluate the fit of the model regardless of its nested nature. Both fit statistics combine model fit with model complexity and penalize for the number of

covariance parameters estimated, with BIC carrying the larger penalty (Dedrick et al., 2009). As both AIC and BIC are computed from a deviance statistic FIML is the suggested estimation method (McCoach & Black, 2012). The preferred model for both is indicated by a lower number.

Stage One Analysis

In this study stage one of the analysis applies a two level random slope model to answer research question number one – how is national testing policy related to the degree schools incorporate testing into their practices and policies? – and research question number two – does the incorporation of testing into a school structure vary by the school's economic and/or academic composition? Dependent variables during this stage of analysis include all six school accountability variables, school policies that are associated with gaming the system, as well as the school compliance composite. HGLM is conducted for the six school accountability measures and dichotomous gaming the system variables while HLM is used for the extra-curricular activities and the school compliance composite. For each dependent variable six different models will be tested for a total of 66 models. Table 7 breaks down the variables included in each model. See Appendix D for the equations for the stage one models. Model one used NTP to predict school practice. In model two, national level control variables are included. In model three, national level education variables are added, including a level two interaction term of market competition by NTP. In models four and five, school level control variables are included and in the last two models school level equity measures are added to address research question number two: models six (school SES and it cross level interaction with NTP) and model seven (school achievement and its cross level interaction with NTP).

Table 7: Stage 1 Model Specifications

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
National Testing Policy (NTP)	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Economic Development		**X**	**X**	**X**	**X**	**X**	**X**
Education Expenditure		**X**	**X**	**X**	**X**	**X**	**X**
Market Competition		**X**	**X**	**X**	**X**	**X**	**X**
Market Competition X NTP		**X**					
School Type			**X**	**X**	**X**	**X**	**X**
School Location			**X**	**X**	**X**	**X**	**X**
School SES				**X**		**X**	
School SES X NTP						**X**	
School Achievement					**X**		**X**
School Achievement X NTP							**X**

X = Included in model

Stage Two Analysis

Stage two of the analysis used a three level random slope HLM to address research question number three – how does the national testing policy and corresponding policy coupling influence student academic outcomes? The dependent variable during stage two includes both student achievement and non-academic outcomes. Three level HLM models significantly increase the difficulty of interpretation; therefore, the models specified at this stage are limited to essential variables. Table 8 breaks down the models for this stage of the analysis. NTP is included at the national level for all models. Separate models are then run for each school accountability variable, gaming the system policy, and school compliance composite as well as their cross level interaction by NTP, yielding 13 models. The final model includes all school practice variables, in addition to student level control variables.

The Form Sanc/Rew coefficient in the final model is interpreted as the difference in math achievement (or other dependent variable respectively) between students in Form Sanc/Rew systems and Summative systems, taking into account differences in school practice.

Table 8: Stage 2 Model Specification

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
NTP	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Compliance			**X**											**X**
Compliance X NTP			**X**											**X**
School Monitoring				**X**										**X**
School Monitoring X NTP				**X**										
School Comparison					**X**									**X**
School Comparison X NTP					**X**									
Parent-School Comparison						**X**								**X**
Parent-School Comparison X NTP						**X**								
Publicly Posted							**X**							**X**
Publicly Posted X NTP							**X**							
Principal Evaluation								**X**						**X**
Principal Evaluation X NTP								**X**						
Teacher Evaluation									**X**					**X**
Teacher Evaluation X NTP									**X**					
Admission Decision										**X**				**X**
Admission Decision X NTP										**X**				
Transfer Decision											**X**			**X**
Transfer Decision X NTP											**X**			
Standardized Tests												**X**		**X**
Standardized Tests X NTP												**X**		
Extra-curricular													**X**	**X**
Extra-curricular X NTP													**X**	
Gender		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Student SES		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Home Language		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Immigrant Status		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**

X = Included in model.

Software

Stata statistical software, version 12 was used for stage one of this analysis. For models with a continuous dependent variable the xtmixed command was used (Skrondal & Rabe-Hesketh, 2012a). For models with a binary dependent variable the xtmelogit command was used (Skrondal & Rabe-Hesketh, 2012b). Both commands use a maximum likelihood approach for estimation. The maximum likelihood estimation (mle) function in xtmixed provides FIML estimates while the default estimation procedure in xtmelogit is adaptive quadrature. HLM software was used for the three level model used during stage two of the analysis (Raudenbush et al., 2011).

CHAPTER FOUR: RESULTS

This results chapter follows the outline provided in the methodology chapter. As over 150 models were estimated in this project only selected results are provided in graphical or tabular form in this chapter. Additional tables are provided in appendices as indicated. The chapter starts with a review of descriptive statistics. Bivariate relationships are then further investigated for identified descriptive trends. A multinomial regression is provided to evaluate the relationship between national level characteristics and NTP. The presence of significant national level predictors supports their inclusion in the multilevel models. Finally, two and three level random slope HLMs are applied to address the three research questions. Statistical significance is set at $p<.05$ for all analyses.

Descriptive Statistics

Table 6 (seen previously) provides descriptive statistics for all variables for the overall sample as well as decomposed by NTP. The first noticeable trend is the similarity in math achievement scores across categories with all NTP category mean scores falling between 465 and 476. A second general observation is the unique attributes in the No NTP category. Overall the characteristics of countries without a national testing policy appear to differ from the three categories with a NTP. Countries without a NTP are economically more developed, with higher SES schools and students, and above average private school enrollment. Additionally, a greater number of immigrant students and students that speak a language other than the test language are found in No NTP countries. While the anomalies in this category are of interest they are not the focus of this study.

Comparing the Summative, Evaluative, and Form Sanc/Rew categories reveal interesting trends at the national and school level. Countries with testing for accountability systems (Evaluative and Form Sanc/Rew) have a higher proportion of overall school

enrollment attending private schools. Schools in testing for accountability systems are more likely to participate in school level accountability practices (designated with AC on Table 6). Although school monitoring appears to be a ubiquitous practice, commonly implemented across all categories, the proportion of schools that compare school performances, publicly post school level results, and use student test scores to evaluate principals and teachers is higher in testing for accountability countries. Continuity in the school accountability trends is not surprising given the bivariate correlations which indicate small to medium correlations between school accountability variables, ranging from .15 to .57. Furthermore, schools in these countries more often participate in standardized tests and include student achievement scores as a criterion in their student admission and transfer decisions. Prominently absent from these trends are differences in school compliance. Descriptive statistics suggest that schools are approximately equally likely to comply with their NTP regardless of their NTP category.

At the school level, there is a wide range in time students spent weekly in math classes, with the highest category, Form Sanc/Rew, spending over a half hour more on mathematics weekly compared to the lowest category, Evaluative. In the students perception of school climate (designated with CL on table 6) variables, little difference is apparent across NTP categories, however, as categories shift toward greater testing for accountability a small positive trend is found in students belief that their teacher in interested in their well-being. Over 90% of students in all categories agree or strongly agree that school is not a waste of time.

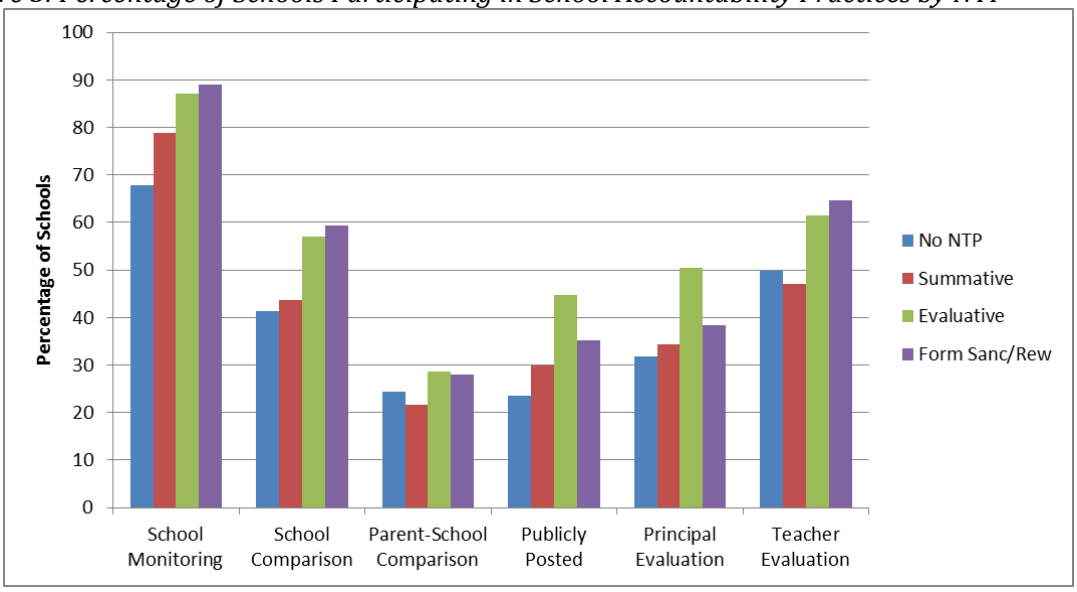
Bivariate Relationships

Based on the descriptive results additional bivariate analyses were completed with NTP and the following variables: school SES, school type, school accountability practices,

admission decision, transfer decision, standardized tests, and time in math. A one way ANOVA found a significant difference in school SES between NTP categories ($F(3,446168)=12772.67$, $p<.001$). A Sidak post hoc test found significant differences between school SES and all bivariate NTP comparisons at the $p<.001$ level. The largest mean differences were between the Form Sanc/Rew category and all other categories, indicating that schools in the Form Sanc/Rew category have significantly lower levels of SES than those without a NTP, or those in Summative or Evaluative categories. Additionally, a one way ANOVA found a significant difference in time spent on mathematics between NTP categories ($F(3,409099)=2309.65$, $p<.001$). A Sidak post hoc test found significant differences in time spent on mathematics between Form Sanc/Rew and all other NTP categories ($p<.001$). The difference in time spent in Evaluative systems and No NTP systems was also significant ($p<.001$). No significant difference was present between Evaluative and Summative categories.

Bivariate analysis between school accountability practices and NTP found a statistically significant relationship between NTP and all school accountability variables: school monitoring ($\chi^2=1.1e+04$, $df=3$, $p<.001$), school comparison ($\chi^2=9.7e+03$, $df=3$, $p<.001$), parent-school comparison ($\chi^2=2.5e+03$, $df=3$, $p<.001$), publicly posted ($\chi^2=9.9e+03$, $df=3$, $p<.001$), principal evaluation ($\chi^2=9.8e+03$, $df=3$, $p<.001$), teacher evaluation ($\chi^2=1.1e+04$, $df=3$, $p<.001$). In all comparisons the highest percentage participation in school accountability practices was found in a testing for accountability category with the lowest percentage in either the No NTP or Summative category (See Figure 5).

Figure 5: Percentage of Schools Participating in School Accountability Practices by NTP



Both the decision to include student achievement in the admission decision (χ^2= 7.0e+03, df=3, p<.001) and the decision to transfer a student (χ^2= 1.0e+04, df=3, p<.001) was significantly related to NTP. Furthermore, there was a statistically significant relationship between whether a school conducts more than two standardized tests during the school year and their NTP (χ^2= 2.2e+03, df=3, p<.001).

Multinomial Regression: Predicting NTP

To investigate whether NTP categories are explained by country characteristics a multinomial regression was conducted regressing national level variables on NTP categories. In this analysis and all that follow Summative is used as the reference group. As the No NTP category is omitted in analyses that include the school compliance variable (schools in countries without a national testing policy cannot comply with a national testing policy), this allows for easy comparison across all analyses. Table 9 displays the results of the multinomial regression and indicate that relative to the Summative category, wealthier countries are more likely to be in the Evaluative or No NTP category. Higher expenditure on education and greater school enrollment in the private sector increases the likelihood that

the country has a Form Sanc/Rew or Evaluative policy, compared to the Summative category.

Table 9: Predicting NTP category by national level variables.

	No NTP	Evaluative	Form Sanc/Rew
n_ed_expend	.832*** (.006)	1.708*** (.007)	1.439*** (.007)
n_develop_c	1.001*** (.001)	1.001*** (.001)	.999*** (.001)
n_comp	.907*** (.001)	1.019*** (.001)	1.038*** (.001)
Constant	.714*** (.026)	.042*** (.001)	.039*** (.001)

Note: Odds Ratios Provided. Standard Errors in Parentheses. Pseudo R-Squared=.074

Between Group Variance

Prior to conducting the HLM analysis an intraclass correlation (ICC) for the final stage 2 analysis in which student achievement acts as the dependent variable was calculated (see Table 10). The ICC suggests that, overall, a greater proportion of the variance in student achievement scores can be attributed to the differences between countries, relative to differences between schools. Roughly 45% of the variation in student achievement scores can be attributed to between student differences, with approximately 25% attributed to differences between schools and 30% attributed to between country differences. As the ICC in all categories is above the minimal threshold of .10 suggested by Lee (2000), multilevel modeling is the appropriate approach.

Table 10: Variation in Math Achievement by Level (ICC)

	Level 1 Variance	Level 2 Variance	Level 3 Variance	Level 1 ICC	Level 2 ICC	Level 3 ICC
PV1Math	4950.189	2720.054	3268.991	0.452517	0.248651	0.298832
PV2Math	4979.408	2723.239	3263.026	0.454091	0.248342	0.297567
PV3Math	4979.778	2713.764	3261.630	0.454560	0.247715	0.297725
PV4Math	4964.114	2712.976	3286.980	0.452762	0.247442	0.299796
PV5Math	4974.265	2704.438	3265.353	0.454518	0.247115	0.298368
Average	**4969.551**	**2714.894**	**3269.196**	**0.453689**	**0.247853**	**0.298458**

Using equation 3, the design effect is calculated from the ICC average on Table 10. The level two design effect of 7.79 [1 + (28.26-1)*.249] and level three design effect of 2015.95 [1 + (6762.58-1)*.298] substantiates the conclusion that multilevel modeling is needed in this study to take into account the clustered design of the data.

First Stage Analysis

Research Question 1: How is national testing policy related to the degree schools incorporate testing into their practices and policies?

After assessing bivariate relationships and ensuring that the data structure is sufficiently clustered to merit the use of multilevel modeling a series of HGLMs were conducted on each school level outcome variable, as outlined on the model specification matrix (see Table 7). The results address research question number one – how is national testing policy related to the degree schools incorporate testing into their practices and policies? – by identifying significant relationships between NTP and school accountability variables and other related school policies (i.e. compliance, admission decision, transfer decision, standardized tests, and number of extra-curricular activities).

School Accountability Practices

To estimate the likelihood of the six school accountability practices by NTP an additive 2 level random intercept HGLM was conducted using the xtmelogit command in Stata statistical package, version 12. For all school accountability variables it is hypothesized that the presence of school accountability practices is more likely in Evaluative and Form Sanc/Rew systems. Table 11 provides estimated odds ratios that schools use student assessments to monitor their yearly progress. Model 1 includes only NTP categories, with Summative used as the reference group. A significant relationship between NTP and school monitoring was not found, however, the likelihood test relative to a single level logistic approach indicates that significant between country variance is present to justify the use of a multilevel model (χ^2=28852.67, df=1, p<.001). Model 2 adds national level control variables and an interaction term (national competition * NTP) that explores whether the relationship between private school enrollment and school monitoring varies by NTP category. The interaction term displaying no significant relationship between private school enrollment (n_comp) and testing for accountability systems (evaluative and form_sanc_rew) and was therefore dropped in later models to ease interpretation. Model 3 adds school level control variables to Model 2. A significant relationship between No NTP systems and school monitoring exists once school type, school location, and national level covariates are controlled for. Specifically, schools in a country that do not have a national testing policy are .238 times as likely as schools in a Summative country to monitor annual school progress through student assessment scores (OR= .238, 95% CI: .075-.738, p<.05). Put another way, schools in No NTP countries are approximately 76% less likely to participate in school monitoring. Model 4 adds school SES to Model 3 while Model 5 adds school mean achievement to Model 3.

Table 11: Odds Ratio of National Testing Policy on School Monitoring

Fixed Effects	Model 1	Model 2	Model 3	Model 4	Model 5
no_ntp	.344 (.203)	.369 (.395)	.238* (.139)	.235* (.137)	.237* (.139)
evaluative	1.218 (.453)	1.645 (.703)	1.827 (.683)	1.826 (.637)	1.825 (.639)
form_sanc_rew	1.506 (.755)	2.131 (1.209)	2.256 (1.023)	2.330 (1.061)	2.231 (1.019)
n_develop_c		.999* (.001)	.999* (.001)	.999* (.001)	.999* (.001)
ed_exp		.682** (.085)	.683** (.083)	.672** (.082)	.679** (.083)
n_comp		.970* (.015)	.974** (.008)	.975** (.008)	.975** (.008)
n_comp*no_ntp		.936 (.118)			
n_comp*evaluative		1.010 (.022)			
n_comp* form_sanc_rew		1.001 (.020)			
private			.990 (.014)	.943*** (.014)	.970* (.014)
sc_loc_town			1.304*** (.016)	1.278*** (.016)	1.295*** (.016)
sc_loc_city			1.278*** (.017)	1.224*** (.017)	1.261*** (.017)
sc_loc_largecity			1.283*** (.023)	1.223*** (.023)	1.263*** (.023)
sc_ses				1.107*** (.011)	
sc_math_c					1.001*** (.001)
Intercept (_cons)	7.295 *** (1.781)	59.973*** (38.174)	48.745*** (31.168)	54.555*** (35.040)	50.389*** (32.452)
Random Effects					
Residual (sd_cons)	1.313 (.115)	1.060 (.102)	1.065 (.102)	1.070 (.103)	1.073 (.103)
Model Fit Statistics					
Deviance	346113.8	299274.8	297323.8	297116.0	297268.3
AIC	346123.8	299290.2	297347.9	297142.0	297294.3
BIC	346178.9	299377.3	297478.4	297283.3	297435.7

Note: Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

The relationship between No NTP and school monitoring remains significant after controlling for school level SES and achievement. In addition to the relationship between No NTP and school monitoring, the complete models suggest that school monitoring is more likely to be practiced in lower income countries (OR=.999, 95% CI: .99994-.99999, p<.05), countries that spend a lower percentage of their GDP on education (OR=.676, 95% CI: .532-.859, p<.01), countries with less private school competition/enrollment (OR=.975, 95% CI: .958-.991, p<.01), public schools (OR=.962, 95% CI: .935-.989, p<.01), schools in more urban areas, higher achieving school (OR=1.001, 95% CI: 1.0001-1.0010, p<.001), and wealthier schools (OR=1.107, 95% CI: 1.086-1.128, p<.001). A likelihood ratio test to compare model fit across nested models revealed that Model 4 best fit the data. This is supported by AIC (297142.0) and BIC (297283.3) statistics where lower value indicates better fit.

Additive models were completed for the additional school accountability variables with output from model 4 of each variable used to create Table 12. General trends in school accountability practices can be deciphered by looking at coefficient patterns across each school accountability variable. Controlling for the other covariates included on the table, five of the six school accountability practices are more likely in Evaluative and/or Form Sanc/Rew categories, relative to Summative, with the largest statistically significant difference found in whether a school publicly posts their achievement results: schools in Form Sanc/Rew systems are nearly 4 times as likely to publicly post their results than schools in Evaluative systems (OR=3.989, 95% CI: 1.452-10.961, p<.01). Other trends include an interesting relationship between education expenditure and school accountability practices with greater spending on education decreasing the likelihood that a school practices the identified accountability policy. Although expenditure is not a significant

factor in estimating whether a school provides comparable school level results directly to parents, a one percent increase in education expenditure as a percentage of GDP leads to a 26% (school comparison and principal evaluation) to 41% (teacher evaluation) decrease in the likelihood the school practices the identified accountability policy.

Table 12: Odds Ratio of National Testing Policy on School Accountability Practices

Fixed Effects	sc_monitor	sc_comp	p_sc_comp	posted	pr_eval	t_eval
no_ntp	.235* (.137)	.453 (.269)	1.042 (.608)	.358 (.237)	.640 (.380)	.470 (.317)
evaluative	1.826 (.637)	2.39* (.834)	2.089* (.727)	3.083** (1.219)	3.012** (1.068)	2.355* (.947)
form_sanc_rew	2.330 (1.061)	2.484* (1.150)	1.385 (.629)	3.989** (2.057)	1.282 (.592)	1.299 (.681)
n_develop_c	.999* (.001)	.999 (.001)	.999 (.001)	1.001 (.001)	.999** (.001)	.999*** (.001)
ed_exp	.672** (.082)	.739* (.092)	.827 (.101)	.669** (.092)	.739* (.091)	.589*** (.083)
n_comp	.975** (.008)	.962*** (.008)	.964*** (.008)	.965*** (.009)	.981* (.009)	.979* (.010)
private	.943*** (.014)	1.103*** (.014)	.983 (.014)	.694*** (.009)	.870*** (.011)	1.677*** (.022)
sc_loc_town	1.277*** (.016)	1.065*** (.011)	.923*** (.010)	1.190*** (.012)	1.186*** (.012)	1.033** (.011)
sc_loc_city	1.224*** (.017)	.935*** (.010)	.890*** (.011)	1.063*** (.012)	1.228*** (.014)	1.050*** (.012)
sc_loc_largecity	1.222*** (.023)	.981 (.014)	.999 (.016)	.982 (.015)	1.390*** (.021)	1.141*** (.018)
sc_ses	1.107*** (.011)	1.030*** (.008)	.893*** (.007)	1.293*** (.010)	.897*** (.007)	.994 (.008)
Intercept (_cons)	54.555*** (35.040)	5.170* (3.373)	.823 (.527)	2.924 (2.125)	1.944 (1.265)	14.926*** (11.030)
Random Effects						
Residual (sd_cons)	1.070 (.103)	1.089 (.104)	1.067 (.102)	1.213 (.116)	1.086 (.104)	1.234 (.118)
Model Fit Statistics						
Deviance	297116	440651.8	367127.4	406662.0	410879.8	395506.0
AIC	297142.0	440677.8	367153.4	406688.0	410885.7	395532.0
BIC	297283.3	440819.1	367294.8	406829.3	411027.1	395673.3

Note: Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

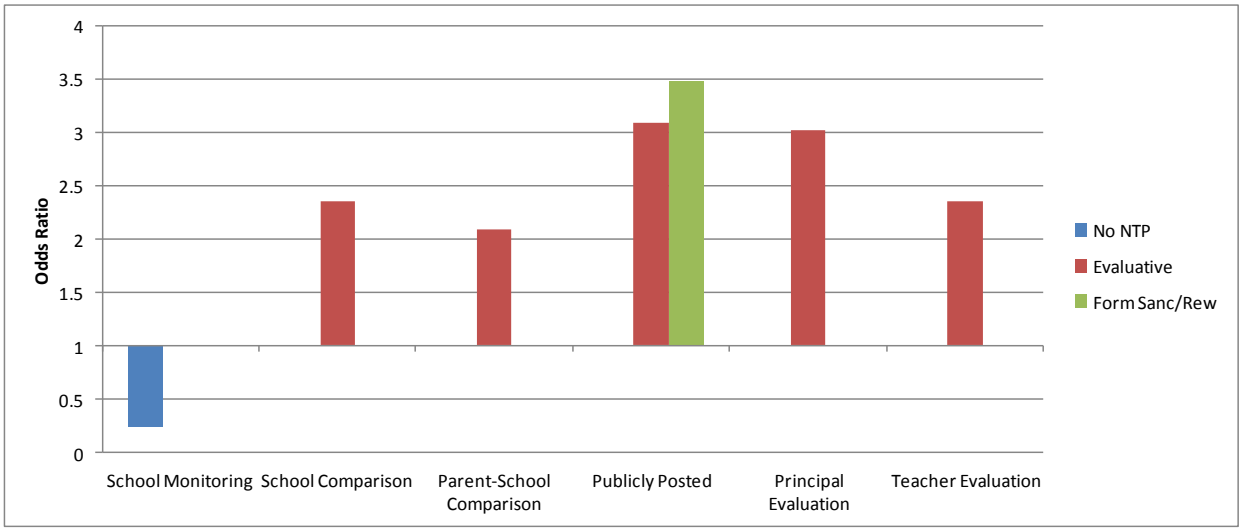
Relative to public schools, private schools strongly encourage certain practices (school comparison and teacher evaluation) while avoiding others (school monitoring, publicly posted, and teacher evaluation). With the exception of providing parents comparable school achievement scores, school accountability practices are, relative to small town, more likely to be found in urban areas. Finally, school economic status can either decrease or increase the likelihood of the given school accountability practice. Higher income schools are more likely to participate in school monitoring (OR=1.107, 95% CI: 1.086-1.128, p<.001), compare their school achievement to other schools (OR=1.030, 95% CI: 1.015-1.046, p<.001) and publicly post their results (OR=1.293, 95% CI: 1.274-1.313, p<.001). In contrast, lower income schools are more likely to provide parents with comparable data (OR=.893, 95% CI: .879-.907, p<.001) and use student test scores to evaluate the principal (OR=.897, 95% CI: .884-.911, p<.001). Noticeably absent from these trends was the practice of evaluating teachers based on student achievement, which is not significantly related to school SES. Finally, the lower deviance statistic, AIC and BIC indicate that the included variables best fit the data for school monitoring.

The high correlation (0.633) between school mean SES and school mean achievement leads to a similar pattern of results when school SES is replaced with school achievement (see Appendix E for complete table)[2]. Figure 6 illustrates significant odds ratios for national testing policy categories after taking into consideration all national and school level control variables, as well as school achievement. The x-axis is set at OR=1 to signify the reference category, Evaluative. Schools in Form Sanc/Rew testing countries are 3.526 times more likely to publicly post their results than schools in a Summative country (CI: 1.224-10.158, p<.05). This practice, however, is the only school accountability variable

[2] Unless otherwise specified, coefficients provided in the remaining sections are taken from model 5; controlling for national level covariates, school type, school location, and school achievement.

where a significant difference in likelihood is found between Form Sanc/Rew and Summative categories. By comparison, in Evaluative countries schools are more likely to practice every type of school accountability, except for school monitoring. The relative lack of a relationship between school monitoring and NTP categories may be due to the previously mentioned ubiquitous implementation of school monitoring within all NTP categories.

Figure 6: Odds Ratio of School Accountability Practices by NTP Category, Evaluative is Reference Category (i.e. OR=1.00)



Note: All odds ratios shown are significant at the p<.05 level.

School Compliance

School compliance, the rough coupling variable that measures the degree a school complies with the national testing policy they are immersed in, is not significantly related to NTP category. As the No NTP category by definition has no testing policy to comply with it is omitted from the model. In general, more compliant schools are found in public schools (β=-.158, p<.001), lower income schools (β=-.049, p<.001) and lower performing schools (β=-.0001, p<.001).

Other School Level Policies

Past research on shaping the testing pool suggests that embedded school policies based on student achievement, such as using student scores to make student admission and transfer decisions, will be more common in testing for accountability countries. To test this hypothesis, models 4 and 5 where estimated with admission and transfer decision as the dependent variables. Results indicate no significant relationship between NTP categories and such decisions. However, both embedded policies are more likely in private schools, high SES schools, higher achieving schools, and schools in countries that spend less on their education (see Appendix F for results).

Prior research on teaching to the test suggests schools experiencing higher levels of school based accountability are more likely to engage in test practice and reduce other non-test related activities. However, no support was found for either of these hypotheses (see Appendix G for results). NTP category is not significantly related to the number of standardized tests the school took, nor is NTP category significantly related to the amount of extra-curricular activities offered at the school.

To summarize, results suggest that national testing policy has a significant relationship with school incorporation of testing into their practices and policies. Specifically, after controlling for national and school level covariates, schools in the two testing for accountability categories are more likely to participate in five of the six school accountability practices, compared to those in Summative systems. Schools, however, are just as likely to comply with their national testing policy regardless of which NTP category they fall under. Finally, the embedded policies basing admission decisions or transfer decisions on test scores, as well as the number of extra-curricular activities and the

proclivity for a school to take more than two standardized tests per year are not statistically related to national testing policy.

Research Question 2: Does the incorporation of testing into a school structure vary by the school's economic and/or academic composition?

In testing for accountability systems low achieving or low SES schools may be more likely to incorporate testing into their school structure as they struggle to avoid punishment, capture rewards, or attract the student enrollment necessary for their survival. To test the hypothesis of differential implementation within NTP categories and identify equity concerns a cross-level interaction term (sc_ses*NTP category or sc_math_c*NTP category) was added to model 4 or model 5, as appropriate. See equation 12 for an example of the two level random slope model used in this analysis.

$$School\ Practice_{jk} = (\delta_{00} + \delta_{01}\text{NTP} + \delta_{02}\text{n\_develop\_c} + \delta_{03}\text{n\_ed\_expend} + \delta_{04}\text{n\_comp}) + \gamma_{1k}\text{sc\_private} + \gamma_{2k}\text{sc\_loc\_} + \gamma_{3k}\text{sc\_ses} + \gamma_{4k}(\text{sc}_{\text{ses}}\text{x NTP}) + \nu_{0k} + \mu_{jk}$$

(Equation 12)

Table 14 provides the odds ratios of the cross level interaction term with each column representing a different dependent variable. Equity concerns in countries that employ a testing for accountability system are apparent on the table. More advantaged school (i.e. high achievement and/or high SES) are more likely to participate in school accountability practices in countries without a national testing policy while less advantaged (i.e. low achievement and/or low SES) schools in testing for accountability countries are overall more likely to implement all six school accountability practices. The two noticeable exceptions to this trend are school monitoring and educator evaluation practices in Form Sanc/Rew countries where higher achieving (sc_monitor: OR= 1.001, CI: 1.0001-1.0011,

p<.05; t_eval: OR= 1.001, CI: 1.0002-1.0010, p<.01) and higher income schools (sc_monitor: OR= 1.058, CI: 1.015-1.103, p<.01; pr_eval: OR= 1.160, CI: 1.125-1.200, p<.001) are more likely to participate.

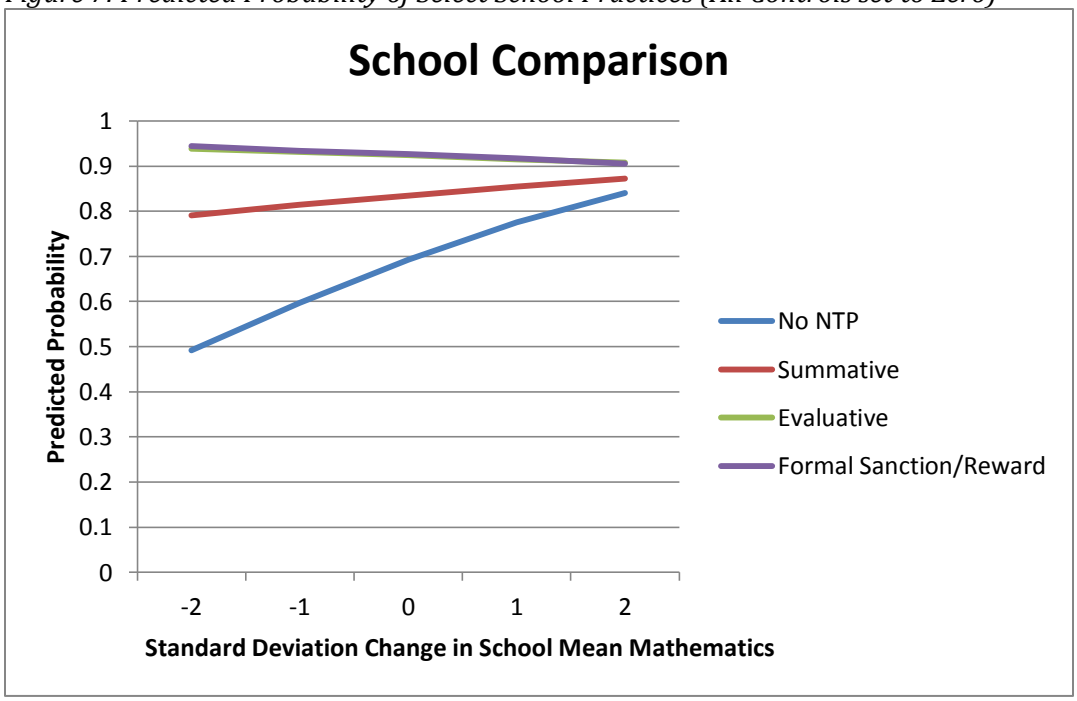
Table 14: Interaction Between School SES and NTP on School Accountability Practices

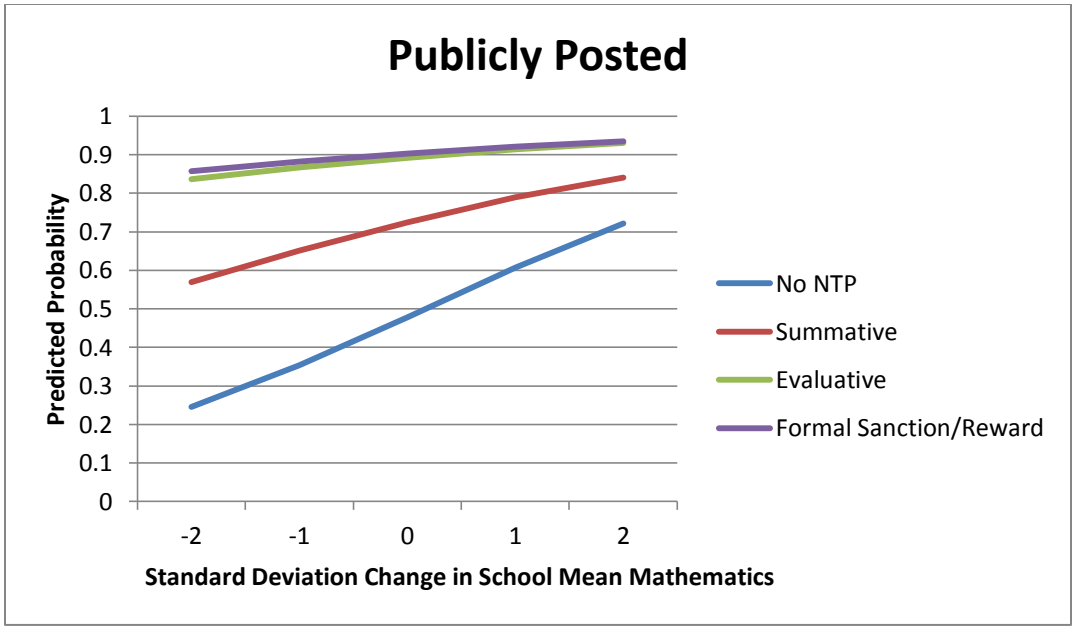
SES Interaction	sc_monitor	sc_comp	p_sc_comp	posted	pr_eval	t_eval
sc_ses	1.114*** (.015)	1.186*** (.013)	1.009 (.013)	1.293*** (.010)	.839*** (.010)	1.020 (.012)
sc_ses*no_ntp	1.403*** (.047)	1.624*** (.052)	1.494*** (.050)	1.523*** (.019)	1.258*** (.041)	1.270*** (.042)
sc_ses*evaluative	.768*** (.018)	.752*** (.012)	.871*** (.016)	.888*** (.015)	1.025 (.017)	.936*** (.016)
sc_ses* form_sanc_rew	1.058*** (.023)	.767*** (.012)	.731*** (.012)	.676*** (.011)	1.160*** (.018)	.937*** (.016)

Note: Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

To better illustrate the equity trends found in Table 14, Figure 7 graphs the predicted probability of school comparison and publicly posted by national testing policy. The graphs in Figure 7 control for all parameters included in the model, setting them all equal to zero. Two trends are illustrated in these graphs. First, the general convergence of probability across NTP categories in schools two standard deviations above the school SES mean suggests that high achieving schools participate in similar practices regardless of their policy context. Second, the greater variation present in low achieving schools across NTP category largely results from the increased likelihood that lower income schools in testing for accountability categories participate in school accountability practices, relative to low income schools in other categories and higher income schools in the given category.

Figure 7: Predicted Probability of Select School Practices (All Controls set to Zero)





The coefficient for the SES interaction term for compliance and the remaining school practice variables are found in Table 15. As both compliance and extra-curricular activities are continuous dependent variables the interpretation of the coefficient differs from the other dependent variables, all dichotomous. For example, in Summative systems a one standard deviation increase in school mean SES corresponds with an increase in extra-curricular opportunities available ($\beta = 1.232$, $p<.001$). The lack of a significant interaction effect

between NTP and extra-curricular activities suggests that the relationship between SES and extra-curricular activities does not significantly differ by NTP category. Higher income schools in Evaluative systems are more likely to participate in more than two standardized tests a year and base admission decisions on student test scores, however, lower SES schools in Evaluative systems are more likely to employ a selective policy.

Table 15: Interaction Between School SES and NTP on Select School Practices

SES Interaction	**admin**	**transfer**	**s_tests**	**ex_act**	**compliance**
sc_ses	1.084*** (.021)	2.307*** (.029)	.859*** (.010)	1.232*** (.151)	-.104 (.056)
sc_ses*no_ntp	2.553*** (.102)	1.452*** (.054)	1.035 (.033)	-.036 (.511)	NA
sc_ses*evaluative	.902*** (.016)	1.189*** (.025)	1.179*** (.021)	-.314 (.229)	.151 (.085)
sc_ses* form_sanc_rew	1.015 (.018)	.555*** (.010)	1.076*** (.018)	-.358 (.287)	-.019 (.112)

Note: **admin, transfer & s_tests:** Odds ratios provided. **ex_act & compliance:** Unstandardized regression coefficients provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

Interesting trends are also apparent in schools in Form Sanc/Rew systems. Of particular interest is the significant odds ratio of the SES-transfer policy interaction term. Schools in Form Sanc/Rew countries with a mean SES one standard deviation below the grand mean are approximately 45% more likely to transfer low achieving students out of their school (OR= .555, CI: 0.536-0.574, p<.001). Following figure 7 predicted probability was calculated for transfer decision. The results indicate that the practice of removing low achieving students through transfer is less likely to be present in Evaluative and Summative systems. This is especially true in the lowest income schools (minus two standard deviations) as the predicted probability that lowest income schools in Summative and Evaluative systems stands at 28.9% and 14.9% respectively. The low predicted probability contrasts sharply with that of Form Sanc/Rew systems which sits at 61.3% for the lowest

income schools. The results suggest that schools in Form Sanc/Rew systems are using their transfer policies to shape their student body, selecting out low achieving students.

In summary, within testing for accountability countries school accountability practices are more likely to occur in less advantaged schools. Additionally, less advantaged schools in Evaluative countries are more likely to make admission decisions based on student achievement, transfer students out that are low achieving, and have more extra-curricular offerings. Less advantaged schools in Form Sanc/Rew countries are more likely to include student test score as a factor in admission and transfer decisions and have more extra-curricular offerings. The differences in transfer decision across school achievement and SES level provides some evidence low achieving schools in Form Sanc/Rew countries are shaping the testing pool to ensure their survival. Relative to similar schools in Evaluative countries, schools in Form Sanc/Rew countries are more likely to transfer a student out of the school due to low academic achievement.

Second Stage Analysis

Research Question 3: How does the national testing policy and corresponding policy coupling influence student outcomes?

To predict student outcomes by NTP category a three level HLM is used. Due the substantial computing power required in three level modeling the following analysis was completed using HLM software, version 6. The No NTP category is omitted from this stage of the analysis due to the extensive use of the compliance variable which captures the degree a school follows their national testing policy. Results from this section address research question three - How does the national testing policy and corresponding policy coupling influence student outcomes? Student outcomes, used as dependent variables in this analysis, include: math achievement, time spent in math, grade repetition, and all four student

reported school climate variables. The series of models outlined in Table 8 was run for each dependent variable.

Student Math Achievement

Table 16 illustrates all models predicting student math achievement. Results reported are adjusted average plausible values, corrected using Rubin's (1987) rule to provide appropriate point estimates and standard errors. Model 1 indicates that testing for accountability systems are significantly related to student achievement, in the absence of control variables. The relationship between Form Sanc/Rew systems and math achievement remains significant after controlling for student level demographic variables (Model 2). Model 3 asks the question, does the degree a school complies with their national testing policy change the relationship between national testing policy and student achievement? When this rough coupling variable is added to the model the coefficient for Evaluative and Form Sanc/Rew is interpreted as the relationship of testing for accountability policy in schools with a zero compliance score with student achievement. The non-significant results of NTP category in model 3 and the overall positive significant result of the cross level interaction with the Evaluative system (evaluative*compliance, $\beta=2.974$, $p<.01$) suggests that in the Evaluative system greater compliance with the national testing policy may play an important role in student math achievement. Models 4 through 13 add school practice variables, used as dependent variables in the earlier analysis, and their corresponding interaction with NTP category. Across these models the relationship between NTP and student achievement tend to be significant, indicating that, when controlling for the given school practice, students in Evaluative or Form Sanc/Rew categories score approximately 34 to 42 points above their peers in Summative systems.

Among the school practice variables, five significant, positive relationships with student achievement are apparent: publicly posted (β=10.205, $p<.001$), teacher evaluation (β=3.432, $p<.05$), admission decision (β=19.131, $p<.001$), transfer decision (β=19.880, $p<.001$) and extra-curricular activities (β=4.666, $p<.001$). One cannot conclude, however, that school practices unanimously have a positive impact on student achievement, as the corresponding interaction term for each school practice suggests heterogeneous effects of school practices on student achievement. For example, the positive association of publicly posting school level results is partially mediated in schools in Evaluative countries by the negative interaction between publicly posted and Evaluative (β=-6.315, $p<.05$). Negative interaction terms, indicating a relative or overall negative relationship of certain practices in schools in testing for accountability countries, are also found in schools that use test scores to compare their school with others (form*sc_comp, β=-8.917, $p<.05$) and schools that provide comparable school results to their parents (form*p_sc_comp, β=-8.121, $p<.05$).

Unsurprisingly, both variables that shape school composition by selecting in high achieving students and out low achieving students are significantly related to student achievement. Students in schools that admit or transfer students based in part on their student achievement have math scores approximately 20 points higher than those in schools that do not submit to that practice. The interaction term with NTP category is not significant. Extra-curricular activities are also positive and significantly related to student achievement. Student achievement scores increase 4.666 points for every additional extra-curricular activity offered at their school with no significant differences found in the NTP interaction term.

Table 16: Estimating Student Math Achievement from NTP and School Practice

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	38.460*	34.069	24.921	37.557*	37.412*	34.047*	35.408*	36.405*	37.499*	36.811	37.100*	35.984*	38.797*	27.742
	(18.619)	(17.321)	(17.694)	(18.266)	(16.598)	(17.038)	(17.425)	(17.003)	(17.473)	(18.734)	(18.009)	(17.496)	(18.435)	(18.379)
Form. San/Rewards	37.258*	38.104*	35.459	40.274*	42.833*	39.843*	38.373*	38.242*	39.573*	33.980	41.932*	40.287*	31.280*	26.301
	(17.901)	(18.332)	(18.927)	(19.913)	(17.918)	(17.739)	(17.866)	(16.869)	(16.989)	(17.111)	(16.403)	(18.670)	(15.102)	(15.398)
*Eval*Inter*			2.974**	-4.141	-6.159	0.215	-6.315*	-4.760	-5.091	-2.298	6.244	-2.599	-1.134	2.704
			(0.936)	(5.208)	(3.490)	(3.174)	(2.497)	(2.955)	(3.005)	(6.818)	(8.166)	(4.584)	(0.937)	(1.575)
*Form.*Inter*			0.865	-2.663	-8.917*	-8.121*	-2.999	-0.634	-2.413	5.037	-12.696	-6.895	0.351	1.699
			(1.104)	(4.006)	(3.495)	(3.104)	(4.938)	(3.218)	(2.960)	(6.545)	(6.882)	(4.745)	(0.981)	(1.892)
Compliance			-1.618*											-2.223*
			(0.692)											(1.067)
sc_monitor				4.107										-3.273
				(3.451)										(3.659)
sc_comp					5.134									-0.803
					(2.709)									(1.808)
p_sc_comp						-2.992								-8.789***
						(1.903)								(0.887)
Posted							10.205***							4.713**
							(1.954)							(1.476)
pr_eval								-1.340						-7.678***
								(1.838)						(1.813)
t_eval									3.432*					-0.747
									(1.561)					(1.468)
Admin										19.131***				14.567***
										(5.065)				(2.902)
Transfer											19.880***			13.796**
											(5.121)			(3.867)
S_tests_b												1.709		-2.009
												(3.657)		(1.680)
ex_act													4.666***	4.154***
													(0.780)	(0.444)
Female		-14.787***	-14.787***	-14.787***	-14.786***	-14.791***	-14.789***	-14.789***	-14.787***	-14.778***	-14.787***	-14.787***	-14.825***	-14.832***
		(1.212)	(1.213)	(1.213)	(1.213)	(1.214)	(1.212)	(1.213)	(1.213)	(1.216)	(1.212)	(1.213)	(1.218)	(1.222)
Immig		-6.522	-6.522	-6.523	-6.527	-6.525	-6.532	-6.517	-6.518	-6.628	-6.528	-6.520	-6.578	-6.678
		(6.148)	(6.149)	(6.150)	(6.148)	(6.148)	(6.148)	(6.148)	(6.148)	(6.139)	(6.148)	(6.147)	(6.154)	(6.147)
lang_other		-9.969***	-9.973***	-9.966***	-9.967***	-9.965***	-9.957***	-9.955***	-9.970***	-9.926***	-10.009***	-9.972***	-9.864***	-9.800**
		(2.738)	(2.736)	(2.739)	(2.738)	(2.737)	(2.739)	(2.738)	(2.736)	(2.750)	(2.744)	(2.737)	(2.746)	(2.769)
Ses		16.529***	16.527***	16.528**	16.529***	16.520***	16.527***	16.530***	16.527***	16.450***	16.473***	16.529***	16.409***	16.299***
		(1.903)	(1.903)	(1.903)	(1.903)	(1.912)	(1.901)	(1.903)	(1.903)	(1.917)	(1.904)	(1.903)	(1.928)	(1.941)
Intercept	442.825	460.448	465.332	456.998	457.716	461.341	457.341	461.011	458.420	448.894	453.483	460.012	426.629	430.946
	(12.852)	(11.901)	(12.778)	(12.850)	(11.546)	(11.633)	(12.077)	(11.605)	(11.823)	(12.867)	(12.074)	(11.555)	(13.072)	(14.438)
Deviance	2928466	2771877	2771865	2771873	2771865	2771852	2771836	2771866	2771743	2771465	2771506	2771729	2771267	2770838

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

In the overall model (model 14) student achievement is a function of NTP category, controlling for school practices and student demographics. With these controls in place, the relationship between NTP and student achievement is no longer significant, suggesting that statistically there is no difference between student scores in Evaluative or Form Sanc/Rew countries and those in Summative countries. The results indicate that, controlling for the included variables, (1) national testing policy is not related to student achievement, (2) greater compliance with the national testing policy is related to lower student achievement, (3) school accountability practices, specifically providing comparable school level results to parents and using student test scores to evaluate principals, are related to lower student achievement, (4) publicly posting school level test scores is related to higher student achievement, (5) practices that shape the student body, namely admission and transfer policies, are related to higher student achievement, and (6) the availability of more extra-curricular activities is related to higher student achievement.

Student Time Spent in Mathematics

To test whether national testing policy narrows the curriculum by shifting resources to tested subjects and away from other areas, Table 17 reports the results of models that use national testing policy, school practices, and student demographics to predict student reported time spent on mathematics. Across all models the amount of time students spend on math in Evaluative systems is not significantly different from the time spent in Summative systems. In comparing Form Sanc/Rew and Summative systems, the difference in time spent on math fluctuates between significant and not significant with students in a Form Sanc/Rew system spending between 35 and 46 minutes more per week on mathematics then their peers in Summative systems. In the final model, students in Form

Sanc/Rew systems spend 44.093 minutes more per week on math ($p<.05$) than students in Summative systems.

Table 17: Estimating Student Reported Time Spent in Math Weekly (Minutes) from NTP and School Practices

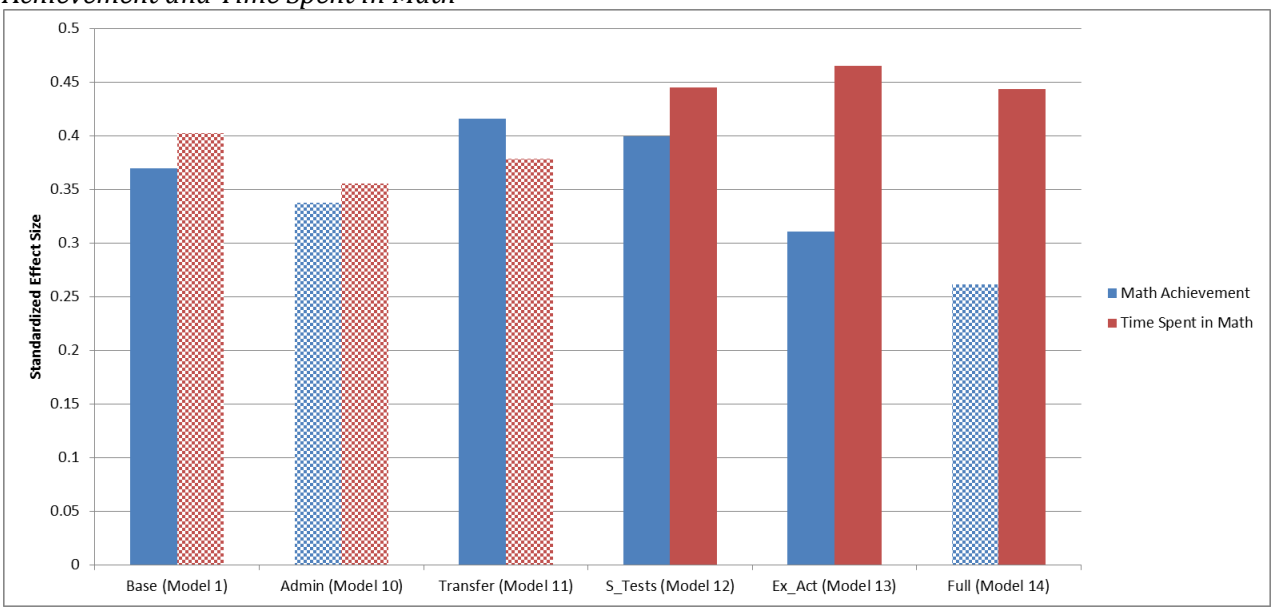
Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	15.539	14.642	10.063	13.808	11.914	15.557	15.169	12.959	14.227	11.146	11.049	17.239	16.559	13.326
	(18.475)	(18.633)	(17.816)	(19.348)	(13.727)	(18.434)	(18.342)	(18.606)	(18.888)	(17.701)	(18.901)	(18.447)	(18.841)	(18.538)
Form. San/Rewards	40.010	40.089	34.076	41.696*	39.704	40.916*	42.316*	39.281	38.253	35.313	37.565	44.220*	46.223*	44.093*
	(20.143)	(20.295)	(19.158)	(19.853)	(20.943)	(20.017)	(20.042)	(20.327)	(20.002)	(19.195)	(20.168)	(19.932)	(21.359)	(19.739)
*Eval*Inter*			1.510	0.971	4.772	-2.761	-2.914	3.404	0.759	6.488	11.144*	-8.747	-0.306	0.056
			(1.398)	(6.891)	(5.574)	(4.265)	(3.622)	(2.825)	(2.436)	(4.756)	(5.133)	(5.011)	(0.733)	(1.386)
*Form.*Inter*			2.010	-1.927	0.707	-2.068	-6.850*	2.482	3.587	7.545	7.523	-13.483	-0.879	-1.465
			(1.525)	(6.739)	(5.474)	(5.256)	(3.291)	(2.496)	(3.038)	(4.879)	(4.839)	(7.137)	(0.682)	(1.966)
compliance			**-1.272**											**0.378**
			(1.170)											**(1.184)**
sc_monitor				**2.282**										1.138
				(6.369)										(2.850)
sc_comp					**-1.914**									-1.823
					(5.300)									(2.501)
p_sc_comp						**5.587**								2.404
						(2.911)								(1.834)
posted							**5.328**							1.499
							(2.904)							(1.963)
pr_eval								**1.204**						1.416
								(1.464)						(1.678)
t_eval									**3.126**					3.554*
									(1.883)					(1.400)
admin										**-3.114**				0.691
										(4.426)				(1.955)
transfer											**-8.936**			-4.476*
											(4.674)			(2.272)
S_tests_b												**11.398***		4.099
												(4.725)		(3.012)
ex_act													**1.186^**	0.751**
													(0.661)	(0.278)
female		-1.900	-1.188	-1.190	-1.191	-1.183	-1.193	-1.188	-1.188	-1.192	-1.188	-1.185	-1.207	-1.948
		(0.891)	(0.891)	(0.890)	(0.890)	(0.892)	(0.890)	(0.890)	(0.890)	(0.890)	(0.891)	(0.893)	(0.889)	(0.891)
immig		-0.567	-0.567	-0.576	-0.562	-0.560	-0.578	-0.572	-0.569	-0.572	-0.559	-0.566	-0.591	-0.593
		(1.141)	(1.141)	(1.140)	(1.141)	(1.141)	(1.142)	(1.140)	(1.141)	(1.137)	(1.142)	(1.142)	(1.144)	(1.141)
lang_other		-0.686	-0.694	-0.684	-0.686	-0.708	-0.683	-0.706	-0.693	-0.698	-0.682	-0.709	-0.647	-0.668
		(1.265)	(1.264)	(1.263)	(1.265)	(1.268)	(1.262)	(1.265)	(1.266)	(1.262)	(1.268)	(1.273)	(1.263)	(1.269)
ses		1.801***	1.799***	1.801***	1.803***	1.799***	1.792***	1.802***	1.795***	1.796***	1.828***	1.800***	1.748***	1.776***
		(0.422)	(0.423)	(0.423)	(0.423)	(0.424)	(0.420)	(0.424)	(0.423)	(0.425)	(0.427)	(0.422)	(0.415)	(0.424)
Intercept	-14.610	-12.972	-9.142	-14.896	-11.950	-14.633	-14.597	-13.476	-14.823	-11.089	-9.817	-16.355	-21.607	-23.488
	(12.300)	(12.393)	(12.104)	(13.536)	(19.385)	(12.617)	(12.069)	(12.604)	(12.589)	(11.353)	(12.439)	(12.189)	(13.915)	(12.550)
Deviance	2758701	2657068	2657062	2657065	2657064	2657042	2657059	2657059	2657055	2657059	2657045	2657037	2657047	2657012

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

Additional school level factors related to time spent on mathematics include: whether teacher evaluations are based on student test scores (β=3.554, $p<.05$), whether a school transfers out low achieving students (β=-4.476, $p<.05$), and the amount of extra-curricular activities offered at the school (β=0.751, $p<.01$). The latter result suggests that time spent in the testing subject and availability of extra-curricular activities are not mutually exclusive.

To examine the differences between the role of national testing policy on student math achievement and the role of national testing policy on student time spent on math, Figure 8 plots the standardized coefficients of models 10 through 14 from tables 16 and 17. These models are chosen because the included variables account for much of the explanatory power of national testing policy on student achievement. Figure 8 illustrates how the inclusion of the admin variable fully mediates the significance of Form Sanc/Rew NTP on both math achievement and time spent in mathematics. Additionally, the magnitude of the Form Sanc/Rew category on time spent in math increases once extra-curricular activities are controlled. At the same time the effect size and significance of the Form Sanc/Rew category on math achievement decreases. In the last model the differences between student achievement scores in Form Sanc/Rew systems is not significantly different from scores in the Summative system. The differences in time spent in math, however, remain significant after controlling for national, school, and student level variables, suggesting that, not only is there no significant difference in math achievement between systems but, students in Form Sanc/Rew systems are spending additional time in math classes and producing the same achievement results.

Figure 8: Standardized Effect Size of Formal Sanction/Reward Testing Policy on Math Achievement and Time Spent in Math



Note: Solid color signifies significant effect (p<.05).

Grade Repetition and Student Perception of School Climate

As final models on the other student non-achievement outcomes do not indicate a direct significant relationship with the NTP category, only the final model (model 14) for the remaining dependent variables (grade repetition and all student reported school climate variables) are provided here (see Appendix H for complete models). What follows is a brief synopsis of significant school level results from Table 18. School practices that are associated with an increased likelihood of a student having ever repeated a grade include providing parents with comparable school scores (OR=1.249, CI: 1.125-1.388, p<.001) and administering more than two standardized tests per year (OR=1.138, CI: 1.022-1.266, p<.05). Factors that reduce the likelihood a student has repeated a grade include: publicly posting school level results (OR=0.888, CI: 0.834-0.947, p<.001), having a selective admissions (OR=0.739, CI: 0.642-0.851, p<.001) and transfer (OR=0.777, CI: 0.672-0.898,

p<.01) policy, and have more extra-curricular activities available (OR=0.930, CI: 0.905-0.955, p<.001).

Table 18: Odds Ratio of National Testing Policy and School Practices on Non-achievement Student Outcomes

Variable	Grade Repetition	CL: Prepared for Life	CL: Not a Waste	CL: Teachers are Interested	CL: Teachers Treat Fairly
Evaluative	0.698	1.001	0.813	0.945	0.872
	(0.361)	(0.183)	(0.156)	(0.176)	(0.1128)
Form. San/Rewards	1.619	0.915	1.122	1.048	0.723
	(0.416)	(0.264)	(0.207)	(0.206)	(0.166)
*Eval*Inter*	0.988	1.048*	1.000	1.066***	1.060**
	(0.038)	(0.023)	(0.026)	(0.015)	(0.022)
*Form.*Inter*	1.059	1.045*	0.978	1.060***	1.047
	(0.035)	(0.020)	(0.028)	(0.015)	(0.024)
compliance	1.017	0.965**	0.997	0.970**	0.965**
	(0.017)	(0.012)	(0.017)	(0.009)	(0.011)
sc_monitor	0.999	0.952*	1.028	0.982	0.994
	(0.048)	(0.024)	(0.030)	(0.020)	(0.024)
sc_comp	0.993	0.960*	0.996	0.986	0.980
	(0.053)	(0.017)	(0.021)	(0.017)	(0.018)
p_sc_comp	1.249***	0.960	0.966	0.975	0.969
	(0.054)	(0.029)	(0.023)	(0.016)	(0.029)
posted	0.888***	1.029	1.066**	0.977	0.991
	(0.032)	(0.016)	(0.023)	(0.017)	(0.017)
pr_eval	0.973	0.994	1.010	0.981	1.001
	(0.040)	(0.020)	(0.023)	(0.016)	(0.021)
t_eval	1.049	0.967*	0.986	0.985	0.957
	(0.031)	(0.014)	(0.025)	(0.016)	(0.027)
admin	0.739***	1.082*	1.174***	0.993	1.064**
	(0.072)	(0.031)	(0.025)	(0.016)	(0.021)
transfer	0.777**	0.998	1.089**	0.961	0.983
	(0.074)	(0.023)	(0.026)	(0.022)	(0.019)
S_tests_b	1.138*	0.987	0.998	1.057**	1.031
	(0.055)	(0.018)	(0.023)	(0.016)	(0.017)
ex_act	0.930***	1.020**	1.027***	0.997	1.011**
	(0.014)	(0.006)	(0.005)	(0.003)	(0.004)
female	0.701***	1.304***	2.050***	1.226***	1.303***
	(0.025)	(0.024)	(0.037)	(0.030)	(0.03)
immig	1.245	0.978	1.247**	1.048	1.022
	(0.163)	(0.043)	(0.079)	(0.028)	(0.042)
lang_other	1.213**	0.879**	0.896	0.954	0.916*
	(0.062)	(0.043)	(0.060)	(0.032)	(0.041)
ses	0.758***	1.092***	1.195***	1.034	1.024
	(0.051)	(0.015)	(0.022)	(0.019)	(0.014)
Intercept	0.218	2.456	6.992	2.417	4.006
	(0.289)	(0.149)	(0.122)	(0.143)	(0.109)
Deviance	63270	268798	20048	277616	227150

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

Across the four student reported school climate variables, selective admission policy and extra-curricular activities are positively associated with better outcomes. Students in schools with these practices feel better prepared for life (admin, OR=1.082, CI: 1.050-1.115, p<.05; ex_act, OR=1.020, CI: 1.015-1.026, p<.01), feel that school is not a waste of time (admin, OR=1.174, CI: 1.118-1.233, p<.001; ex_act, OR=1.027, CI: 1.017-1.036, p<.001),

and feel that their teachers treat them fairly (admin, OR=1.064, CI: 1.022-1.109, p<.01; ex_act, OR=1.011, CI: 1.004-1.018, p<.01). Additionally, the use of school level accountability practices increase the likelihood that students feel less prepared for life (sc_monitor, OR=0.952, CI: 0.909-0.998, p<.05; sc_comp, OR=0.960, CI: 0.930-0.992, p<.05; t_eval, OR=0.967, CI: 0.940-0.994, p<.05). Finally, the significance of the compliance variable and its interaction terms in three of the four school climate dependent variables indicate that compliance with the national testing policy is more important in testing for accountability countries. This result, however, should be interpreted cautiously as many of the school accountability practices that were used to calculate compliance have a negative effect on student perceptions. This may indicate that schools that comply with their national testing policy are more aligned, in general, with their national policies which may have beneficial effects not captured in this study.

CHAPTER FIVE: DISCUSSION AND CONCLUSSION

Standardized testing for accountability is becoming a normative global practice. Increasingly countries are disseminating results at the school level and using student test scores as a barometer from which to apply formal sanctions or rewards. Through the categorization of the 2009 PISA participants into national testing policy categories, results from this study can guide policymakers looking to use accountability to improve educational efficiency, equity, and excellence. Specifically, this analysis questions whether the optimistic belief that testing for accountability is associated with higher levels of student achievement is a robust finding or a fabrication of selective admission and transfer policies. In using market and behavioral approaches to shape educator behavior, it is important to recognize that survival instincts may contrast with effective instructional practices and furthermore, that these behavioral choices are due to an environment that incentivizes practices that shape the testing pool and allocates more time to tested subjects. This discussion chapter concludes the study in four parts. First, the studies aims and unique contributions to the literature are reviewed. Second, the main findings are summarized, situating the results of the three primary research questions within the previous literature. Third, policy recommendations are suggested from the results. Finally, limitations to the study and areas for future research are discussed.

Study Aims and Contribution to the Field

This study expands the literature on accountability by attempting four conceptual advances. First, previous research has historically identified accountability through simple binary measures, identifying whether or not accountability is present in the system. In this study I recognize the multiple aims of testing and try to map a spectrum of accountability policies onto the identified aims. Once testing for accountability is identified it is further

disaggregated into Evaluative and Form Sanc/Rew policies. The distinction between policies that rely on market tactics (Evaluative) to shape educator practice and that which is based on behaviorism is important as it can guide educator interventions. Furthermore, the preponderance of past studies focus on tests that hold students accountable (see Au 2007; Lipman 2004). While this is important it is not the emphasis of this study. Instead testing for accountability, which places accountability firmly on educators, is central to this analysis as it has the ability to swiftly transform school and classroom practices and policies by shaping educator behavior.

Secondly, this study contributes to the field by using multiple dependent variables to capture more holistically the effect of national testing policies. While secondary data and quantitative measures alone may not be able to fully illustrate the cultural impressions left by testing policies the use of school level practices and student non-academic outcomes expands the dominate work which focuses on student achievement scores (de Wolf & Janssens 2007; Lee 2008). The inclusion of student non-academic outcomes is especially important as past research on equity indicates that schools under pressure may create a hostile climate by preferencing some groups of students over others (Hursh, 2007; Teelken, 1999). Additionally, repetition rates may be more prominent in testing for accountability systems attempt to remove potential low scoring students from the testing pool (Hursh, 2005; Kornhaber, 2004a; Kornhaber, 2004b).

Thirdly, past research has been dominated by single country studies (Figlio & Loeb 2011). Woessmann (2007) questions whether the reliance on single country samples are appropriate given the minimal within country variance associated with policy enactment. This study, however, questioned the assumption of minimal within country variance. Education is often characterized by devolution with local policy enactors having high levels

of autonomy. Essentially, schools and teachers are often given guidance on the ends but not the means. This leads to a variety of practices as educators attempt to do what they can to teach their specific set of students. Methodologically this study created a proxy variable to capture the loose coupling between the announced policy and the enacted policy. Although it provides an imperfect measure this school compliance variable is a first attempt at identifying heterogeneous within country practices as well as which schools are more likely to comply with national policy. Conceptually, the compliance variable is necessary in a three level design. Regressing student level outcomes on national level policies (when students are clustered into classrooms) raises questions on whether schools follow the national level policy. Results that indicate low levels of compliance cloud the interpretation of the national level predictor variable. Although the school compliance variable played a marginal role in this study and produced largely non-significant results future study testing other ways to capture policy coupling are necessary.

In addition to examining within country variance, this study added to the previous research by pooling a large cross-national sample into four national testing policy categories. The large sample provided by the 2009 PISA allowed for the categorization of countries by differences in national testing policy while maintaining some within category heterogeneity (economically, regionally). Although trend data indicate substantial between category differences, the diversity of countries within each category adds to the strength of the national testing policy argument by illustrating how the potential effects of national testing policy largely transcend cultural differences. In the end this approach results in a rich description of what testing for accountability looks like at the school and student level.

Main Findings

Incorporating these conceptual advances into the analysis yielded interesting insights on how national testing policy relates to school practices and student outcomes that, at times, contrasted with previous research. This section summarizes the main findings of this study, by research question.

Research Question 1: How is national testing policy related to the degree schools incorporate testing into their practices and policies?

National level policies are associated with differing educator behavior. In exploring how national testing policy relates to school practices and policies, at least three main conclusions are apparent. First, testing for accountability systems, especially Evaluative systems, are associated with a greater likelihood that schools will use test scores for school level accountability practices. Schools in Evaluative systems are approximately two times more likely to participate in all school accountability practices, such as publicly posting their school level test results or using test scores to evaluate teachers, relative to Summative systems. Second, the relationship between NTP category and gaming the system practices yielded mixed results. No significant relationship was found between NTP and shaping the testing pool variables (i.e. selective admission and transfer policies). Similarly, unlike past research that found schools in accountability systems were more likely to teach to the test (Au, 2007; Cuban, 2007; Nicols & Berliner, 2007), there was no association between the number of standardized tests a school participated in and NTP in the adjusted analysis.

Third, there is not a significant association between NTP and the availability of extra-curricular activities. This result contrasts with previous research that finds schools in accountability systems often divert resources away from non-tested activities (Ben Jaafar & Anderson, 2007; Figlio & Loeb, 2011). One potential explanation for this inconsistency may

be that as schools shift resources to tested subjects (i.e. mathematics), non-tested subjects, such as art and band, have their within school time reduced or eliminated. To make up for this lack of within school opportunity, schools that are able are forced to offer auxiliary opportunities, outside of normal school time.

Research Question 2: Does the incorporation of testing into a school structure vary by the school's economic and/or academic composition?

The incorporation of schooling for accountability practices and policies that game the system vary by the schools socio-economic status. Three primary findings concerning the equitable adoption of school level practices can be drawn from this study. First, schools with a higher mean SES are more likely to participate in publicly posting test results, use test results to monitor yearly progress, use test results to compare their school with others, have selective admission and transfer policies, and offer more extra-curricular activities. Second, interaction terms for Formal Sanction/Reward system indicate that less advantaged schools in this category are significantly more likely to participate in key school accountability practices. Lower income schools in this system are at least 26% more likely to provide parents with comparable school level test results or publicly post their results, compared to low income schools in Summative systems. Third, low income schools in Formal Sanction/Reward systems are more likely to transfer out low achieving students. Low income schools, under immense pressure in Formal Sanction/Reward systems, appear to use selective transfer policies to shape their testing pool, thus increasing their mean achievement score and ensuring their survival. Upon graphing predictive probabilities it appears that practices in more advantaged school are largely homogenous, regardless of NTP category. However, the negative gradient commonly found in the interaction term with Form Sanc/Rew suggests that schools at greatest risk of punitive measures are more likely to

incorporate school accountability practices and policies that can artificially inflate their school mean achievement.

Research Question 3: How does the national testing policy and corresponding policy coupling influence student outcomes?

National testing policies and school accountability practices have little beneficial effects on student academic and non-academic outcomes. The suggestion that testing for accountability systems will increase educational excellence by leveraging educator behavior is not supported in this study. The main findings on student outcomes are reviewed below. First, after controlling for school accountability practices and other school policies related to gaming the system, NTP is not associated with student math achievement. Results indicate selective admission and transfer decisions are related to higher student math achievement. This is not surprising, as schools that adopt this policy only admit or retain students that are higher achieving. Once accounting for admission policy, it becomes clear that there is no significant difference in math achievement in schools that do not shape their testing pool across NTP categories. The impact of selective admission policies on student math achievement confirms the practice of 'cream skimming' in testing for accountability systems (Gerwitz et al., 1995; West, Pennell & Noden, 1998). Although selective admission policies are not more likely to occur in Evaluative or Form Sanc/Rew systems, the mediating effect of admission policies on student achievement in the later system suggests that its influence may be greater in Form Sanc/Rew schools.

Second, publicly posting school level test results was the only school accountability practice directly related to student math achievement. Furthermore, negative interaction terms with the Form Sanc/Rew category indicated that using test scores to compare school results and providing comparable school aggregated test scores directly to parents may be detrimental to student achievement for schools in a Form Sanc/Rew system. Although

schools in Evaluative systems are more likely to participate in nearly all school accountability practices this increased proclivity does not result in a significant gain in student achievement.

Third, students in Form Sanc/Rew systems spend over 40 minutes more per week in mathematics than their peers in Summative systems. This finding supports prior work that found schools in testing for accountability system transfer instructional time to tested subjects, such as mathematics (McNeil & Valenzuela, 2001; Rentner et al., 2006). The contrast in comparing the final models for student math achievement and student time spent in math is important and suggests that the promise of greater efficiency, which accompanies testing for accountability policies, is not being met. The incongruence between increased instructional time and improved achievement has been found in past studies including Falk (1996), McNeil (2000), and Supovitz (2009), and may indicate that the increased instruction is dominated by superfluous adjustments that fail to promote deeper improvements in student understanding.

Fourth, NTP categories are not directly related to student repetition. Across all NTP categories students that take more standardized tests are more likely to have repeated a grade while students in schools that offer more extra-curricular activities are less likely to have repeated a grade. Additionally, students in schools that select in or select out their student body are less likely to have repeated a grade.

Policy Recommendations

Findings from this study indicate that the equity concerns of testing for accountability systems outweigh the non-significant achievement results, a relationship more pronounced in Form Sanc/Rew systems. As policymakers are evaluating the appropriate approach to testing and accountability they need to reconsider using policies that

incentivize educator behavior that has negative consequences for the less advantaged students and is not related to gains in achievement. The following seven policy recommendations, drawn from past research and the results of this study, outline suggestions and alterations that are needed to ensure excellence in education does not come at the expense of equity.

Policy Recommendation #1: Test results should be disseminated at the regional or national level.

Providing school level results to the public and using school level results as the foundation for sanctions and rewards leads to undesirable behavior modification from educators attempting to ensure their professional survival. School level aggregation and dissemination incentivize school accountability practices that are not related to student achievement in testing for accountability countries and encourages gaming the system policies which purposefully omit less advantaged students. Disseminating regional or national results provide countries with system level data that can be disaggregated by gender, ethnicity, urbanicity, etc., allowing policymakers to explore equity trends. Dissemination at the macro level shifts responsibility to the state, reinforcing the position of education as a public good. With dissemination at the macro level, student scores would not be withheld from parents, but instead provided relative to the regional or national mean for the comparable demographic.

Policy Recommendation #2: Conduct sample based, not census based, tests.

Conducting annual, census based tests uses important instructional time for test implementation and corresponding test prep. Taking a census of students provides an opportunity for school aggregation which can lead to problems associated with school level dissemination. In contrast, conducting tests based on a sample of students reduces the

overall pressure associated with a testing culture, reducing the likelihood that educators game the system by narrowing the curriculum or teaching to the test. A representative sample can be used for system level analysis, providing insight into the state of education and important equity concerns. Additionally, the testing subjects can be rotated or done at random. For example, Estonia tests language arts and mathematics during every round but rotates one other subject from a pool that includes different science and social studies topics as well as foreign language, while the ministry of education in the Slovak Republic decides on tested subjects in the months leading up to the testing period. Rotating tested subjects provide more information for policymakers while not overloading students by testing every subject during the testing period. Testing more subjects over a larger time frame encourages a more holistic educational experience by placing value on a greater number of subjects.

Policy Recommendation #3: Mandate open access admission policies for public schools.

Selective admission and transfer policies are not limited to private schools. In this study over 50% of public schools used student achievement as a condition of admittance and over 30% transferred out low achieving students. As demonstrated in the results section, the later policy is more likely to be present in less advantaged schools in Form Sanc/Rew systems. Selective admission policies limit access to education by forbidding entrance from students deemed academically undesirable. This concentrates poorer performing students in struggling schools, creating a bifurcated system where the advantaged schools extend their advantage. To ensure equity, governments must eliminate achievement as a selection criterion for public schools.

Policy Recommendation #4: Strengthen the position of educators as professionals.

Implementing policy that strengthens the position of educators as professionals will lead to greater trust within schools and between schools and the community. This can lead to

more effective within-class assessment that provides the necessary immediate feedback to inform instruction. Multiple steps can be taken to improve the standing of educators as professionals. First, entrance to the profession needs to be more selective and be marked by higher expectations. Communities need to have faith that the pre-service education and training of educators properly prepares them for work in schools. Expectations for post-secondary coursework need to be increased and pedagogical skill should be emphasized. Gradual release of responsibility in the classroom to pre-service teachers should be encouraged over rushed training programs that only last a few months or programs where teachers learn on the job. Second, teacher capacity in assessment must be increased. Capacity should be focused on ongoing, in-class assessments that provide immediate feedback for educators, allowing for swift adjustments in pedagogical strategy and overcoming the longer feedback loop that plagues large-scale assessments (Chapman & Snyder, 2000; Kornhaber, 2004a; Paris & McEvoy, 2000; Supovitz, 2009). It is also important to recognize that assessments encompass much more than testing (Darling-Hammond, 1994; Froese-German, 2001, Kornhaber, 2004a). Good teachers are able to monitor the progress of their students and provide evidence of their proficiency through daily assessments, including class discussion, debates, projects, group work, skits, labs, homework, and presentations. Third, educators should be incorporated into the policy decision making process. If we want to know what education in a community is truly like, it is important to include those individuals on the front line as stakeholders. Educators are professionals with valuable insights into their practice and must be seen as a resource. Improving the position of educator as a profession will improve community confidence in schools. Seeing educators as professionals reduces the need for nonstop monitoring, instead

educators are trusted because they have a unique set of skills and know what is best for the student's education.

Policy Recommendation #5: Promote practices that foster cooperation not competition.

Education is a cooperative profession that works best when educators share resources and best practices for the benefits of all students. Testing for accountability policies destroy this cooperative environment by pitting individual educators and schools against one another. Evaluative and Form Sanc/Rew systems create an environment where there are winners and losers. In such an environment it is in the schools best interest to not disclose effective practices with competitors. Competitive policies place the blame solely on the individual, stymieing collective responsibility and discarding important societal factors that contribute to educational outcomes (segregation, income inequality, poor funding, etc.).

Policy Recommendation #6: Accountability systems should have multiple indicators and encourage educator agency and capacity building.

Accountability systems should be focused on providing a comprehensive measure of educational quality that goes beyond mere student achievement. To capture student understanding along with more challenging to measure attributes such as school climate and student well-being, feedback from parents and students should be incorporated into school and educator evaluation. Additionally, accountability components should be both external and internal in nature, drawing on a sense of collective responsibility where educators have the ability and the respect to provide quality peer review. This type of accountability system is obviously reliant on the prior policy recommendations involving teacher professionalism and a cooperative environment, illustrating the complexity and interdependency of education policy. In addition to increasing educator agency by including them as stakeholders in the accountability process, outcomes of accountability should be focused on capacity building to improve instruction. The punitive consequences often attached to accountability do little to

enhance the educational quality in schools that need it most. Instead of removing funding, thereby harming students, struggling schools should be provided with the resources necessary to improve their practice. Finally, reward systems need to be restructured. Pay for performance schemes are often unsustainable and generally involve reducing or holding constant the base teacher salary which diminishes the position of educator as professional. Instead the base salary of teachers should be increased and rewards linked to accountability should be used to improve instructional practice by providing educators with additional preparation time, funding for a lesson study, or the opportunity to match schools with differing strengths in a mutually beneficial relationship.

Policy Recommendation #7: Support holistic education.

Results on extra-curricular activities as well as past research speak to the importance of a holistic, well-rounded education for student achievement and well-being. Testing for accountability policies undermine holistic education by encouraging curriculum narrowing and teaching to the test. Experiences with subjects such as music, the arts, and foreign language do not only provide students with a well-rounded knowledge base but are related indirectly to achievement gains in the commonly tested subjects of mathematics and language arts. Funding and resources need to be allocated to such subjects to ensure student access to a breath of educational experiences.

Limitations and Future Research

The data selection and categorization process for this multi-level comparative study are a first attempt to create national level variables based on accountability policies focused on educators. Results of this study should be cautiously interpreted as preliminary due to its novel nature. Replication of this study with an expanded set of countries over multiple time points is necessary to solidify trends in national testing policy categories. Limitations

specific to data categorization, generalizability of results, and validity are discussed in this section.

Creating categorizes is often a subjective process. As much as a researcher selects seemingly objective criteria for differentiating categories, multiple assumptions are required to uniquely classify actors into one single category. It is important to consider whether national testing policy categories are mutually exclusive. This analysis assumed that countries may have multiple national or regional tests at the elementary or lower secondary level and that these tests fall into different categories. When multiple tests were present for a given country the country was categorized into the highest or most intense accountability category. This categorization decision was based on an assumption that the presence of any test for accountability can shape educator behavior, regardless of the presence of other tests that may be classified as Summative. Additionally, limiting the tests used for coding to those administered at the elementary or lower secondary level assumes that tests students were exposed to prior to the age of 15 is of greatest importance when exploring the influence of national testing policy on 15 year olds. It may be the case, however, that only tests during the survey year directly influence the school practices and student outcomes. This is unlikely given the ability of tests and testing practices to become immersed in the larger school culture. It could also be the case that tests at the upper secondary level influence school practices at the lower secondary level. This is more likely is schools that combine lower secondary and upper secondary grades, placing students from both levels in the same physical space and educators in the same professional culture. The relationship between the presence of tests at the upper secondary level and school practices and student outcomes at the lower secondary level is an empirical question that deserves further study. Of particular interest may be the interaction between Evaluative and Form Sanc/Rew testing policies and

the traditionally identified "high-stakes" tests that often come during the transition from upper secondary to tertiary education.

National testing policy categories are also sensitive to the included countries. For example, the Form Sanc/Rew category may look different if the U.S. is excluded from the analysis. The Form Sanc/Rew category is especially prone to changes in variances, as it includes a relatively small number of countries making each country responsible for a larger percentage of the variance. This limitation is dictated by the participating countries in the 2009 PISA. As testing for accountability becomes the expected policy practice of the 21st century more countries will shift into the Form Sanc/Rew category, allowing for more robust findings and clearer interpretations. The relatively small country sample within the Form Sanc/Rew category also contributes to the large standard errors in the three level model. These large standard errors make the interpretation of the regression coefficient challenging. Future studies that include more countries and multiple time points are needed to examine whether the findings of this study are a product of category sensitivity and whether the non-significant findings remain robust when a larger sample is included in the Form Sanc/Rew category. One potential solution to the category sample is to collapse Evaluative and Form Sanc/Rew into a single category. This was not done in this study to maintain the distinction between market and behavioral incentives in shaping educator behavior and to capture the more direct punitive consequences in Form Sanc/Rew systems.

Categorization concerns also affect the generalizability of results. Specifically, the 2009 PISA data is predominately a sample of information from developed countries. The results and policy suggestions outlined above may not be appropriate for developing countries that are more likely to be working on system infrastructure and fulfilling universal access to education requirements. Nevertheless, testing for accountability is spreading into

these developing regions and future case studies should be conducted to explore whether trend data presented here remains consistent in a developing country context. Additionally, case studies in countries such as Bangladesh and Uganda can provide important insights to how educators respond to different incentives in a more resource poor environment.

The use of secondary data to complete a cross-national large scale analysis leads to a reliance on proxies to measure key concepts, bringing up questions of validity. To maintain a comparative framework this analysis includes more simplified measures that are available across the entire sample. For example, although percentage of private school enrollment has been used in previous research as a proxy for competition between schools it is unable to capture whether a particular school feels direct competition. Jensen, Weidmann, and Farmer (2013) raise additional questions on whether schools are competing over academics or other consumer demanded attributes and whether consumers (parents) consider only within sector (public or private) competition or evaluate schools across sectors. These important questions about competition cannot be clearly addressed using secondary data from over 60 countries; instead these questions must be addressed through national level studies or smaller comparative analyses that collect information directly from parents regarding choice decisions and educators regarding the pressure they feel from competition.

In addition to the suggested case studies to explore contextual differences and provide more precise measures of competition, choice, and educator practice, future studies should expand their analysis of the testing culture beyond the boundaries of the school. The transformation toward testing for accountability globally creates an atmosphere where test scores are equated with education quality and parents are increasingly expected to make decisions based on school aggregated results. This testing culture can affect other aspects of society including: whether the nation-state is seen as legitimate internationally; the state's

role in education; funding for education, including the amount and sources, both domestically and internationally; the education-labor force link and the importance of educational credentials; how educators are trained; and how the actors in education (student, teacher, parent, etc.) are labeled by society. The permeation of the testing culture throughout society can have a widespread impact on education and its role in social and economic development. This study is the first step in exploring this larger testing culture.

REFERENCES

Apple, M. (1999). Rhetorical reforms: Markets, standards, and inequality. *Current Issues in Comparative Education*, 1(2), 6-17.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258–267.

Baker, D. (2014). *The Schooled Society: The Educational Transformation of Global Culture.* Stanford, CA: Stanford University Press.

Baker, D. & LeTendre, G. (2005). *National Differences, Global Similarities: World Culture and the Future of Schooling*. Stanford, CA: Stanford University Press.

Baker, D., Thorne, S., Blair, C. & Gamson, D. (2006). Cognition, culture, and institutions: Affinities within social construction of reality. Unpublished paper. University Park, PA: The Pennsylvania State University.

Baker, D., & Wiseman, A. W. (Eds.). (2005). *Global Trends in Educational Policy* (Vol. 6, International Perspectives on Education and Society Series). Oxford, UK: Elsevier Science.

Ball, S. J. (1995). Parents, schools and markets: The repositioning of youth in United Kingdom education. *Young*, 3(3), 68-79.

Ball, S.J. (1993). Education markets, choice and social class: The markets as a class strategy in the UK and the USA. *British Journal of Sociology of Education*, 14(1), 3-18.

Barry, C., Horst, J., Finney, S., Brown, A. & Kopp, J. (2010). Do examinees have similar test taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, 10, 342-363.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *The Journal of Political Economy*, 70(5), 9-49.

Belfield, C. R., & Levin, H. M. (2002). The effects of competition between schools on educational outcomes: A review for the United States. *Review of Educational Research*, 72(2), 279-341.

Benavot, A., & Tanner, E. (2007). The growth of national learning assessments in the world, 1995-2006. Background paper for the Education for All Global Monitoring Report 2008: Education for All by 2015: Will we make it.

Ben Jaafar, S. & Anderson, S. (2007). Policy trends and tensions in accountability for educational management and services in Canada. *Alberta Journal of Educational Research*, 53(2), 205-225.

Bishop, J., Mane, F., Bishop, M., & Moriarity, J. (2001). The role of end-of-course exams and minimal competency exams in standards-based reforms. *Brookings Papers in Educational Policy*. Washington, DC: Brookings Institution.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Education Research Journal*, 42(2), 231-268.

Bruns, B., Filmer, D. & Patrinos, H. (2011). *Making Schools Work: New Evidence on Accountability Reforms*. Washington, DC: The World Bank.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods* (First Edition). Newbury Park, CA: Sage Publications.

Butland, D. (2008). *Testing Times: Global Trends in Marketisation of Public Education through Accountability Testing.* Marrickville NSW: NSW Teachers Federation.

Carl, J. (1994). Parental choice as national policy in England and the United States. *Comparative Education Review,* 38(3), 294-322.

Carney, S., Rappleye, J. & Silova, I. (2012). Between faith and science: World Culture theory and comparative education. *Comparative Education Review*, 56(3), 366-393.

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis,* 24(4), 305–31.

Certo, J. (2006). Beginning teacher concerns in an accountability-based testing environment. *Journal of Research in Childhood Education*, 20(4), 331-349.

Chapman, D. & Snyder, C. (2000). Can high stakes national testing improve instruction: Re -examining conventional wisdom. *International Journal of Educational Development*, 20, 457-474.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93, 1045-1057.

Chubb, J. & Moe, T. (1990). *Politics, Markets, and America's Schools*. Washington, D.C: Brookings Institution Press.

Clotfelter, C., Ladd, H., Vigdor, J. & Diaz, R. (2004). Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? *J. Policy Anal. Manag.* 23(2), 251–271.

Cohen, D.K. & Ball, D.L. (1999). Instruction, capacity, and improvement. In *Consortium for Policy Research in Education (CPRE) Report Series RR-43* (pp. 1-41). Philadelphia, PA: University of Pennsylvania Graduate School of Education.

Cohen, D. K., & Rosenberg, B. H. (1977). Functions and fantasies: Understanding schools in capitalist America. *History of Education Quarterly*, *17*(2), 113-137.

Cronin, J., Kingsbury, G.G., McCall, M.S., & Bowe, B. (2005). *The Impact of the No Child Left Behind Act on Student Achievement and Growth.* Evanston, IL: Northwest Evaluation Association.

Cuban, L. (2007). Hugging in the middle. Teaching in an era of testing and accountability, 1980–2005. *Education Policy Analysis Archive*, 15(1).

Cullen, J. & Reback, R. (2006). Tinkering towards accolades: School gaming under a performance accountability system. In T. Gronberg & D. Jansen (Eds.) *Advances in Applied Microeconomics* (pp. 1-34). London: Emerald Group Publishing Limited.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, 64(1), 5-30.

Datnow, A. & Park, V. (2009). Conceptualizing policy implementation: Large-scale reform in an era of complexity. In G. Sykes, B. Schneider & D. Plank (Eds.) *Handbook on Education Policy Research* (pp. 348-361). New York: Routledge Publishing.

Dedrick, R., Ferron, J., Hess, M., Hogarty, K., Kromrey, J., Lang, T., Niles, J. & Lee, R. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69-102.

Dee, T. & Jacob, B. (2009). The impact of No Child Left Behind on student achievement. *National Bureau of Economic Research Working Paper No. 15531*.

Department of Education (1994). *Our Children's Education: The Updated Parent's Charter*. London: DFE.

De Wolf, I. & Janssens, F. (2007). Effects and side effects of inspection and accountability in education: An overview of empirical studies. *Oxford Review of Education*, 33(3), 379-396.

Diez Roux, A.V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, 56, 588-594.

Dillon, S. (2011). Obama to waive parts of No Child Left Behind. *New York Times* (September, 22, 2011).

Dorn, S. (2007). *Accountability Frankenstein: Understanding and Taming the Monster.* Charlotte, NC: Information Age Publishing.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1).

Drori, G., Meyer, J.M., Ramirez, F.O. & Schofer, E. (Eds.) (2003). *Science in the World Polity: Institutionalization and Globalization*. Stanford, CA: Stanford University Press.

Earl, L. & Torrance, N. (2000). Embedding accountability and improvement into large-scale assessment: What differences does it make? *Peabody Journal of Education*, 75(4), 114-141.

Eckstein, M. & Noah, H. (1993). *Secondary School Examinations*. New Haven, CT: Yale University Press.

Edwards, T. & Whitty, G. (1992). Parental choice and educational reform in Britain and the United States. *British Journal of Educational Studies*, 40(2), 101-117.

Eurydice (2009a). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Education, Audiovisual and Culture Executive Agency, European Commission.

Eurydice (2009b). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results – Country Level Reports: Hungary*. Education, Audiovisual and Culture Executive Agency, European Commission.

Falk, B. (1996). Issues in designing a learner-centered assessment system in New York state: Balancing reliability and flexibility, authenticity, and consequential validity. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Feng, L., Figlio, D. & Sass, T. (2009). School accountability and teacher mobility. *University of Wisconsin Working Paper*.

Ferrer, G. (2006). *Educational Assessment Systems in Latin America: Current Practice and Future Challenges*. Washington, D.C.: Partnership for Educational Revitalization in the Americas.

Figlio, D. (2006). Testing, crime, and punishment. *Journal of Public Economics*, 90, 837 -851.

Figlio, D. & Kenny, L. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, 93(9–10), 1069–1077.

Figlio, D. & Loeb, S. (2011). School accountability. *Handbook of the Economics of Education*, 3, 383-421.

Figlio, D. & Lucas, M. (2004). What's in a grade? School report cards and the housing market. *American Economic Review*, 94(3), 591–604.

Fine, M. (1991). *Framing Dropouts: Notes on the Politics of an Urban Public High School*. New York, NY: SUNY Press.

Freidman, M. (1962). *Capitalism and Freedom*. Chicago, IL: University of Chicago Press.

Freidman, M. (1955). The role of government in education. In Solo, R. (ed.) *Economics and the Public Interest*. New Brunswick, NJ: Trustees of Rutgers College in New Jersey.

Friedman, T. (1999). *The Lexus and the Olive Tree.* New York: Random House.

Froese-Germaine, B. (2001). Standardized testing + high-stakes decisions = educational inequity. *Interchange*, 32(2), 111–130.

Frontline (2008). *The Testing Industries Big Four*. From http://www.pbs.org/wgbh/pages/frontline/shows/schools/testing/companies.html.

Gillborn, D. (1996). *Exclusions from School*. London: University of London.

Gillborn, D. & Youdell, D. (2000). *Rationing Education: Policy, Practice, Reform and Equity.* Buckingham, UK: Open University Press.

Gilles, C., Cramer, M. & Hwang, S. (2001). Beginning teacher perception of concerns: A longitudinal look at teacher development. *Action in Teacher Education,* 23(3), 92-98.

Gipps, C. (2003). Educational accountability in England: The role of assessment. Paper presented at the Annual Meeting of the American Education Research Association, New York, NY.

Gipps, C. & Murphy, P. (1994). *A Fair Test?* Philadelphia: Open University Press.

Guo, G. & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology,* 26, 441-462.

Hagerstrom, M. (2006). Basic education. In *Decentralized Service Delivery for the Poor* (pp. 1-47). Washington, DC: The World Bank.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives,* 8(41).

Hanushek, E. & Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management,* 24(2), 297–327.

Hanushek, E. & Raymond, M. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association*, 2(2-3), 406-415.

Harris, D. & Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education,* 112(2), 209-238.

Harvey, D. (2005). *A Brief History of Neoliberalism.* New York, NY: Oxford University Press.

Hill, D. (2006). Class, capital and education in this neoliberal and neoconservative period. *Information for Social Change*, 23, 11–34.

Ho, P. (1964). *The Ladder to Success in Imperial China: Aspects of Social Mobility*. New York: Science Editions.

Hofmann, D. & Gavin, M. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24(5), 623-641.

Hopmann, S.T. (2008). No child, no school, no state left behind: Schooling in the age of accountability. *Journal of Curriculum Studies*, 40(4), 417-456.

Hopstock, P. & Pelczar, M. (2011). *Technical Report and User's Guide for the Program for International Student Assessment (PISA): 2009 Data Files and Database with U.S. Specific Variables* (NCES 2011-025). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mather & M. Schader (Eds.) *Classification, Data Analysis, and Data Highways* (pp. 147-154). New York: Springer Verlag.

Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Educational Research Journal*, 44(3), 493-518.

Hursh, D. (2005). The growth of high-stakes testing in the USA: Accountability, markets and the decline of educational equality. *British Educational Research Journal*, 31(5), 605-622.

Irons, E. J., & Harris, S. (2006). *The Challenges of No Child Left Behind: Understanding the Issues of Excellence, Accountability, and Choice*. Lanham, MD: Rowman and Littlefield Education.

Jacob, B.A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, 89(5–6), 761–796.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, 23(2), 99–121.

Jacob, B.A. & Levitt, S.D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 843–878.

Janssens, F. & Visscher, A. (2004). Towards a school report card for primary education. *Educational Studies*, 81, 371-383.

Jensen, B. Weidmann, B. & Farmer, J. (2013). *The Myth of Markets in School Education*. Grattan Institute.

Jepperson, R. L. (2002). The development and application of sociological neoinstitutionalism. *New Directions in Contemporary Sociological Theory*, 229-266.

Jolly, R. (2003). Human development and neoliberalism: Paradigms compared. In S. Fukuda Parr & A. Shiva Kumar (Eds.) *Readings in Human Development* (pp. 82-92). New York, NY: Oxford University Press.

Jones, B. & Egley, R. (2006). Looking through different lenses: Teachers' and administrators' views of accountability. *Phi Delta Kappan*, 87(10), 767-771.

Joshi, D. & Smith, W. (2012). Education and inequality: Implications of the World Bank's Education Strategy 2020. In A. Wiseman & C. Collins (Eds.), *Education Strategy in the Developing World: Revising the World Bank's Education Policy* (pp. 173-202). United Kingdom: Emerald Publishing.

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.

Kamens, D. H., Meyer, J. W., & Benavot, A. (1996). Worldwide patterns in academic secondary education curricula. *Comparative Education Review*, 40(2), 116-138.

Kim, K., Kim, G., Kim, S., Kim, J. et al. (2010). *OECD Review of Assessment and Evaluation Frameworks for Improving School Outcomes: Country Background Report for Korea.* Paris: OECD.

Kleinman, M., West, A., & Sparkes, J. (1998). *Investing in Employability: Theories of Business and Government in Transition to Work.* Commissioned by BT. London: LSE.

Koretz, D. (2008). Test-based educational accountability. Research evidence and implications. *Zeitschrift für Pädagogik*, 54(6), 777-790.

Kornhaber, M. L. (2004a). Assessment, standards, and equity. *Handbook of Research on Multicultural Education*, 2, 91-109.

Kornhaber, M.L. (2004b). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy*, 18(1), 45-70.

Kreft, I. G. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30, 1-21.

Kremelberg, D. (2010). *Practical Statistics: A Quick and Easy Guide to IBM SPSS Statistics, Stata, and Other Statistical Software.* Thousand Oaks, CA: Sage Publications.

Krull, J. & MacKinnon, D. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36(2), 249-277.

Lashway, L. (2001). *The New Standards and Accountability: Will Rewards and Sanctions Motivate America's Schools to Peak Performance?* Washington, DC: ERIC Clearinghouse of Educational Management.

Lavy, V. (2007). Using performance-based pay to improve the quality of teachers. *Future Child* 17 (1), 87–109.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, 78(3), 608-644.

Lee, V. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, 35(2), 125-141.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., & Jocelyn, L. (2004). *International Outcomes of Learning in Mathematics Literacy and Problem Solving: PISA 2003 Results from the U.S. Perspective*. National Center for Education Statistics. Washington, DC: US Government Printing Press.

Levin, H. (1992). Market approaches to education: Vouchers and school choice. *Economics of Education Review*, 11(4), 279-285.

Lewin, T., & Medina, J. (2003, July 31). To cut failure schools shed students. *The New York Times*, p. Al.

Lincove, J. (2009). Are markets good for girls? The World Bank and neoliberal education reforms in developing countries. *The Whitehead Journal of Diplomacy and International Relations*, x(1), 19-35.

Lipman, P. (2004). *High-Stakes Education: Inequality, Globalization, and Urban School Reform*. New York: Routledge.

Luna, C. & Turner, C.L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *Engl. J.*, 91 (1), 79–87.

Maas, C. & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86-92.

Macnab, D. (2004). Hearts, minds, and external supervision of schools: Direction and development. *Educational Review*, 56(1), 53-64.

McCoach, D.B. & Black, A. (2012). Introduction to estimation issues in multilevel modeling. *New Directions for Institutional Research*, 154, 23-39.

McDonnell, L. & Elmore, R. (1987). Getting the job done: Alternative policy instruments. *Educational Evaluation and Policy Analysis*, 9(2), 157-184.

McLaughlin, M. (1991). Test-based accountability as a reform strategy. *Phi Delta Kappan*, November, 248-251.

McNeil, L. (2000). *Contradictions of School Reform: Educational Costs of Standardized Testing*. New York: Routledge.

McNeil, L. & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In G. Orfield & M.L. Kornhaber (Eds.) *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education* (pp. 127-150). New York: Century Foundation Press.

McNeil, M. & Klein, A. (2011). Obama outlines NCLB flexibility: Plan waves cornerstone provisions of law. *Education Week*, 31(5), 1 & 20-22.

Meyer, J. (2005) *Weltkultur: wie die westlichen Prinzipien die Welt durchdringen,* ed. G. Krücken (Frankfurt, Germany: Suhrkamp).

Meyer, J. (1977). The effects of education as an institution. *American Journal of Sociology*, 83(1), 55-77.

Meyer, J., Ramirez, F, Frank, D. & Schofer, E. (2006). Higher education as an institution. Working Paper. Center on Democracy, Development, and the Rule of Law: Stanford University.

Monk, D., Sipple, J., & Killeen, K. (2001, September 10). Adoption and adaptation, New York state school districts' responses to state imposed high school graduation requirements. An eight year retrospective. Working Paper: University of Albany.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, 22, 376−398.

Nichols, S., & Berliner, D. (2007). *Collateral Damage: How High-Stakes Testing Corrupts America's Schools.* Cambridge, MA: Harvard Education Press.

Nitko, A. J., & Brookhart, S. M. (2006). *Educational Assessment of Students* (5th ed.). New York: Prentice Hall.

O'Connell, A. & Reed, S. (2012). Hierarchical data structures, institutional research, and multilevel modeling. *New Directions for Institutional Research*, 154, 5-22.

OECD (2010). PISA 2009 Data Set. Retrieved from https://pisa2009.acer.edu.au/

OECD (2012). *PISA 2009 Technical Report, PISA*, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

Osborne, J. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, and Evaluation*, 7(1), 1-3.

Paris, S., & McEvoy, A. (2000). Harmful and enduring effects of high-stakes testing. *Issues in Education*, 6(1), 145–159.

Peugh, J. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48, 85-112.

Phelps, R. P. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice*, 19(1), 11–21.

Ramirez, F. P. (2003). The global model and national legacies. In K. Anderson-Levitt (Ed.), *Local Meanings, Global Schooling. Anthropology and World Culture Theory* (pp. 239–254). New York: Palgrave Macmillan.

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

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, R. & du Toit, M. (2011). *HLM 7: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International, LLC.

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

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., et al. (2006). *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Center on Education Policy.

Robertson, H. (1999). In Canada – bogus points. *Phi Delta Kappan*, 80(9), 715-716.

Robertson, H. (1998). *No More Teachers No More Books: The Commercialization of Canada's Schools.* Toronto, ON: McClelland & Stewart.

Rosen, L. (2009). Rhetoric and symbolic action in the policy process. In G. Sykes, B. Schneider & D. Plank (Eds.) *Handbook on Education Policy Research* (pp. 267-285). New York: Routledge Publishing.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. N. (2007). *Success Under Pressure?How Low-Performing Schools Respond to Voucher and Accountability Pressure*. Mimeo. Princeton University and University of Florida.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Sahlberg, P. (2010). Rethinking accountability in a knowledge society. *Journal of Educational Change*, 11, 45-61.

Schmidt, W., Houang, R. & Shakrani, S. (2009). *International Lessons about National Standards*. Washington, DC: Thomas B. Fordham Institute.

Schofer, E., Hironaka, A., Frank, D. J., & Longhofer, W. (2012). Sociological institutionalism and world society. *The Wiley-Blackwell Companion to Political Sociology*, 33, 57-69.

Schultz, T. W. (1961). Investment in human capital. *The American Economic Review*, 51(1), 1-17.

Shulman, L. (1983). Autonomy and obligation: The remote control of teaching. In L. Shulman & G. Sykes (Eds.) *Handbook of Teaching and Policy* (pp. 484-505). New York: Longman.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2012a). *Multilevel and Longitudinal Modeling Using Stata. Volume I: Continuous Responses.* College Station: Stata Press

Skrondal, A., & Rabe-Hesketh, S. (2012b). *Multilevel and Longitudinal Modeling Using Stata. Volume II: Categorical Responses, Counts, and Survival.* College Station: Stata Press

Smeed, J., & Victory, M. (2010). Testing for accountability. *TLN Journal*, 17(3), 28-29.

Smith, W. (2012). Conceptualizing soft power through education. *International and Comparative Education Magazine*, 3. Published online at http://www.icemag.org/2/post/2012/02/conceptualizing-soft-power-through-education.html

Smith, W. & Rowland, J. (2014). Parent trigger laws and the promise of parental voice. *Journal of School Choice*, 8(1), 94-112.

Snijders, T. O. M., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*(3), 342-363.

Springer, M.G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*. doi: 10.1016/j.econedrev.2007.06.004

Stillman, A. & Maychell, K. (1986). *Choosing Schools: Parents, LEAs and the 1980 Education Act.* Windsor: NFER-Nelson.

Stotsky, S. (2000). *What's at Stake in the K12 Standards War?* New York: Peter Lang Publishing.

Subedi, B.R. (2005). A demonstration of the three-level hierarchical generalized linear model applied to educational research. Dissertation. Department of Educational Psychology and Learning Systems: Florida State University.

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, 10, 211-227.

Swadener, B. B., & Lubeck, S. (1995). The social construction of children and families "at risk": An introduction. In *Children and Families "At Promise": Deconstructing the Discourse of Risk* (pp. 1-16). New York, NY: SUNY Press.

Tyack, D. & Cuban, L. (1995). *Tinkering toward Utopia: A Century of Public School Reform*. Cambridge, MA: Harvard University Press.

Volante, L. (2007). Evaluating test-based accountability perspectives: An international perspective. Paper presented at the Association for Educational Assessment – Europe, Stockholm, Sweden.

Von Zastrow, C. & Janc, H. (2004). *Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools*. Washington DC: Council for Basic Education.

Waterreus, I. (2003). *Lessons in Teacher Pay: Studies on Incentives and the Labour Market for Teachers.* Doctoral Thesis: University of Amsterdam.

West, A. & Pennell, H. (2000). Publishing school examination results in England: Incentives and consequences. *Educational Studies*, 26(4), 423-436.

West, A., Pennell, H. & Noden, P. (1998). School admissions: Increasing equity, accountability and transparency. *British Journal of Educational Studies*, 46(2), 188-200.

Wheelock, A., Haney, W., & Bebell, D. (2000). What can student drawings tell us about high-stakes testing in Massachusetts?. *The Teachers College Record*.

Wiggins, A. & Tymms, P. (2000). Dysfunctional effects of public performance indicator systems: A comparison between English and Scottish primary schools. Paper presented at the European Conference on Educational Research, Edinburgh.

William, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, 45(2), 107-122.

Wiseman, A. (2010). The uses of evidence for educational policymaking: Global contexts and international trends. *Review of Research in Education*, 34, 1-24.

Wiseman, A. W., Astiz, M. F., & Baker, D. P. (2013). Comparative education research framed by neo-institutional theory: a review of diverse approaches and conflicting assumptions. *Compare: A Journal of Comparative and International Education*, (ahead-of-print), 1-22.

Wiseman, A. W., Pilton, J., & Lowe, J. C. (2010). International educational governance models and national policy convergence. *International Perspectives on Education and Society*, 12, 3-18.

Woessman, L. (2007). International evidence on school competition, autonomy, and accountability: A review. *Peabody Journal of Education*, 82(2-3), 473-497.

Woessman, L. (2004). The effect heterogeneity of central exams: Evidence from TIMSS, TIMSS-Repeat and PISA. *CESifo Working Paper No. 1330.*

Wong, M., Cook, T.D. & Steiner, P.M. (2009). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series. *Northwestern University Institute for Policy Research Working Paper Series WP-09-11*.

World Bank. (2013). Data retrieved November 2, 2013, from World DataBank: World Development Indicators database.

Yarema, C.H. (2010). Mathematics teachers' views of accountability testing revealed through lesson study. *Mathematics Teacher Education and Development*, 12(1), 3-18.

Ysseldyke, J., Dennison, A. & Nelson, R. (2004). *Large-scale Assessments and Accountability Systems: Positive Consequences for Students with Disabilities.* Minneapolis, MN: National Center on Educational Outcomes.

APPENDIX A: References for Country Categorization

1. Eurydice. (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results.* Brussels: Education, Audiovisual and Culture Executive Agency.
2. Eurydice – [Country] (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results. Country Level Reports.* Brussels: Education, Audiovisual and Culture Executive Agency.
3. Schmidt, W., Houang, R. & Shakrani, S. (2009). *International Lessons about National Standards.* Washington, D.C.: The Thomas Fordham Institute.
4. Schmidt, W., Houang, R. & Shakrani, S. (2009). *International Lessons about National Standards (Tables 3-5).* Washington, D.C.: The Thomas Fordham Institute.
5. Ferrer, G. (2006). *Educational Assessment Systems in Latin America: Current Practice and Future Challenges.* Washington, D.C.: Partnership for Educational Revitalization in the Americas.
6. Personal Communication with Pablo Fraser, Penn State University. Formerly of the Pontifical Catholic University of Chile.
7. OECD (2011), Education at a Glance 2011: OECD Indicators, OECD Publishing. doi: 10.1787/eag-2011-en
8. OECD Reviews of Evaluation and Assessment in Education – [Country]. Available at http://www.oecd-ilibrary.org/education/oecd-reviews-of-evaluation-and-assessment-in-education_22230955
9. National Education Monitoring Project – New Zealand. *About NEMP*. Available at http://nemp.otago.ac.nz/_about.htm
10. Federal Ministry for Education, the Arts, and Culture (2012). *Statistical Guide 2012: Key Facts and Figures about Schools and Adult Education in Austria*. Vienna, Austria: BMUKK.
11. Morris, A. (2011). Student standardised testing: Current practices in OECD Countries and a literature review. *OECD Education Working Papers, No. 65*. OECD Publishing.
12. Bonamino, A. & Sousa, S.Z. (2012). Three generations of assessments of basic education in Brazil: Interfaces with the curriculum in/of the school. *Educacao e Pesquisa*, 38(2), 373-388.
13. Bulgaria. Eurypedia: European Encyclopedia on National Education Systems. Available at https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Bulgaria:Overview
14. UNESCO (2011a). Bulgaria. *World Data on Education*, 7th Edition, 2010/2011.
15. World Bank (2008a). Columbia. The quality of education in Columbia: An analysis and options for a policy agenda. *Report No. 43906-CO*. Human Development Sector Management Unit and Latin America and the Caribbean Regional Office: The World Bank.
16. Finland. Eurypedia: European Encyclopedia on National Education Systems. Available at https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Finland:Overview
17. Iceland. Eurypedia: European Encyclopedia on National Education Systems. Available at https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Iceland:Overview

18. Moos, L. (Ed.) (2013). *Transnational Influences on Values and Practices in Nordic Educational Leadership.* London: Springer Publishing.
19. Leshem, S. & Markovits, Z. (2012). Assessment in schools in Israel: Policies and practices. *SA-eDUC JOURNAL,* 9(2).
20. OECD Economic Surveys [Country] (Various Years). Available at http://www.oecd-ilibrary.org/economics/oecd-economic-surveys_16097513
21. Engel, L., Williams, J. & Feuer, M. (2012). The global context of practice and preaching: Do high scoring countries practice what U.S. discourse preaches? *The George Washington University: Graduate School of Education and Human Development Working Paper 2.3*
22. National Center for Education – Latvia. Available at http://visc.gov.lv/en/exam/statistics.shtml
23. Liechtenstein. Eurypedia: European Encyclopedia on National Education Systems. Available at https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Liechtenstein:Overview
24. OECD (2002). *Reviews of National Polices for Education – Lithuania*. Paris: OECD Publishing.
25. World Bank (2008b). Romania – education and skills for EU integration. World Bank Education Note 73052.
26. Bolotov, V., Lenskaya, E. & Agranovich, M. (n.d.). Improving quality of education in Russia through transforming quality assurance systems. Available at http://www.edu.gov.on.ca/bb4e/russiaEn.pdf
27. Nordic Recognition Network (2005). *The System of Education in Russia.* Copenhagen: Danish Centre for Assessment and Foreign Qualifications.
28. International Qualifications Assessment Services (2008). *International Assessment Guide: For the Assessment of Education from the Former USSR and the Soviet Federation.* Edmonton, Canada: Government of Alberta.
29. Slovak Republic. Eurypedia: European Encyclopedia on National Education Systems. Available at https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Slovakia:Overview
30. Arcia, G., Patrinos, H., Porta, E. & Macdonald, K. (2010). School accountability and autonomy in context: Application of benchmarking indicators in select European countries. The World Bank. Available at http://siteresources.worldbank.org/EDUCATION/Resources/278200-1290520949227/7575842-1339186330807/SAA_Europe_Benchmarking_Note_Feb_7_June2012.pdf
31. OECD (2007). Reviews of National Policies for Education: Basic Education in Turkey. Paris: OECD Publishing.
32. Figlio, D. & Loeb, S. (2011). School accountability. Handbook of the Economics of Education, 3, 383-421.
33. Italian Esame di Stato. Available at http://www.istruzione.it/esame_di_stato/index.html
34. Email Correspondance from Dr. Andris Kangro, University of Latvia to Guillermo Montt, OECD.
35. Email Correspondance with Renata Horvatek, Penn State University. Formerly of University of Zagreb, Croatia.

36. OECD (2010). Reviews of National Policies for Education: Kyrgyz Republic. Paris: OECD Publishing.
37. OECD (2003). Reviews of National Policies for Education: South Eastern Europe. Paris: OECD Publishing.
38. Zellman, G. et al. (2009). *Implementation of the K-12 Education Reform in Qatar's Schools.* Santa Monica, CA: RAND Corporation.
39. Qatar Supreme Education Council. Available at http://www.sec.gov.qa/En/Pages/Home.aspx
40. Dowling, A. (2008). Output measurement in education. Camberwell: Australian Council for Educational Research
41. World Bank (2006). *Thailand Social Monitor: Improving Secondary Education.* Bangkok: The World Bank.
42. Hill, P. (2010). *Asia-Pacific Secondary Education System Review Series No. 1: Examination Systems*. Bangkok: UNESCO-Bangkok.
43. Ministry of Education (2013). *Education Reform for Knowledge Economy: Second Phase. Annual Narrative Report.* The Hashemite Kingdom of Jordan: Development Coordination Unit.
44. Russian Education Aid for Development (2011). Uses of assessment information to support student learning in Jordan. Presentation at the Third READ Conference, Eschborn, Germany (October 25, 2011).
45. International Review of Curriculum and Assessments Framework-Internet Archive (2012). *INCA Comparative Tables*. Available at http://www.nfer.ac.uk/what-we-do/information-and-reviews/inca/INCAcomparativetablesMarch2012.pdf
46. Swiss Education System. Available at https://swisseducation.educa.ch/en
47. World Bank (2012). SABER Country Report – Serbia. The World Bank.
48. UNESCO (2011b). Kazakhstan. *World Data on Education*, 7th Edition, 2010/2011.
49. World Bank (2010). Implementation, completion and results report. *Report No: ICR00001134*. Human Development Sector Unit, South Caucasus Country Unit & Europe and Central Asia Region: The World Bank.
50. Azerbaijan Ministry of Education. Available at http://www.edu.gov.az/view.php?lang=en&menu=78&id=2557
51. Ruban, N. The assessment system in the Ministry of Education of the United Arab Emirates. Powerpoint presentation.
52. Government of the Republic of Trinidad and Tobago. Student Exams. Available at http://www.moe.gov.tt/student_exam_ntest.html
53. African Development Bank (2005). *Republic of Tunisia: Secondary Education Support Project, Phase II. Appraisal Report.* Tunis-Belvedère, Tunisia: African Development Bank.
54. Ministry of Education and Training (2009). *The Development of Education: National Report 2004-2008.* Republic of Tunisia.
55. U.S. Department of Education. Questions and answers on No Child Left Behind. Available at http://www2.ed.gov/nclb/accountability/schools/accountability.html#5
56. UNESCO (2011c). Turkey. World Data on Education, 7th Edition, 2010/2011.
57. UNESCO (2011d). Albania. World Data on Education, 7th Edition, 2010/2011.
58. Manzi, J., Strasser, K., Martin, E.S. & Contreras, D. (2008). *Quality of Education in Chile*. Inter-American Development Bank. Available at *www.iadb.org/res/laresnetwork/files/pr300finaldraft.pdf*

59. OECD (2013), *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices (Volume IV).* Paris: OECD Publishing.

APPENDIX B: Sample Evidence for Testing for Accountability Country Categorization

Evaluative Categorization Rationale

Provided here are key representative references for Evaluative country categorization. For other resources used in categorization see appendix C.

Australia

- Morris (2011, p. 12): "In Australia, NAPLAN test results are published on an individual school basis on the *My School* website, where the public can access performance and other data on schools across Australia."

Azerbaijan

- World Bank (2010): Progress as outlined in World Bank's education sector evaluation, including "progress in making available school statistics and data on student learning" and making the results "available in the Ministry's website" (p. vi).

Belgium (Flemish Community)

- World Bank-Belgium, Flemish Community (2011, p. 21): "The Decree on Transparency in Governance in 2004 allowed for the publication of individual school inspection reports. From 2007 on, these have been published on the Inspectorate's website."

Brazil

- Bonamino & Sousa (2012, p. 379): "The results of the first edition of *Test Brazil* were divulged in July 2006 through the main media and a newsletter made available on the Internet and sent to each participating school. Among other information, this newsletter showed the schools' results on a performance scale and the scores by school in municipal, state and federal systems."

Canada (Ontario)

- Schmidt, Houang & Shakrani (2009, p. 64): "The Ministry issues reports with a profile of students' performance; districts and schools prepare their own reports that include action plans for improving instruction. The results of the assessments are also reported in local newspapers with schools ranked from highest to lowest."

Columbia

- Figlio & Loeb (2011, p. 385): "Countries vary in the degree in which they assess students, and whether they publically report scores at the school level. In Latin America, for instance, scores are publically reported in Brazil, Chile, Columbia, and Mexico (in some regions)."

Denmark

- Eurydice (2009, p. 55): Denmark is one of a few countries which "arrange for central government publication of results for each school, or recommend local publication." Information is provided to the public for the end of compulsory exam.

Estonia

- Eurydice (2009, p. 55): “In Estonia, schools are expected to publicise the aggregated test results of their pupils.”

Iceland

- Eurydice (2009, p. 55): In Iceland school level results are published “with weighted indicators depending on the characteristics of the pupil population or the added value of schools”

Korea, Republic of

- Schmidt, Houang & Shakrani (2009, p. 56): “In 2007, the country decided for the first time to test all students in grades 6, 9, and 10 with plans to release the results at the regional, district, and school levels.”

Lithuania

- Review of Education Policies in Lithuania (2002, p. 107): “The NEC keeps test results by school and can collate them at national, regional and local level to inform education authorities. Since 2000, a student level database is available, as well as full test and item analysis. Every school receives comparative information about its own and national results and all school results are published on the Internet. All information is openly available to the public and the NEC website receives as many as 3 000 visits a day during peak exam periods.”

Netherlands

- Schmidt, Houang & Shakrani (2009, p. 49): “The national primary school leaving examination provides data that compare schools’ average scores with national averages and with those of other schools.”

Norway

- Review of Evaluation and Assessment in Education – Norway (2011, p. 55): “Individual school results are not published on the open part of the School Portal website, but they have been published by the media every year.”

Poland

- OECD (2013): The Matura exam, first administered in 2005, “sent a clear signal to students that their success depended directly on their externally evaluated outcomes, and made it possible to assess teachers and schools on a comparable scale across the whole country” (p. 82).

Qatar

- Zellman et al. (2009, p. xxiv): One of the goals of the 2002 Education for a New Era Reform was to create “a composite index that ranks schools according to student performance and other valued outcomes would inform parent decision making and inspire healthy competition among Independent schools.” To aid between school competition the first round of school report cards was completed in 2006. School report cards apply just to independent schools, however, the goal in Qatar is to convert all ministry schools to independent schools over time.

Romania

- Eurydice (2009, p. 54): "The results of national tests, aggregated for each school, are considered by the central education authorities in the evaluation or auditing of schools." The efficiency of the system, however, is in question as the World Bank Policy Note (2008) suggests that public dissemination of school level results are sporadic at best.

Singapore

- Schmidt, Houang & Shakrani (2009, p. 53): "School results are published in league ranking tables as a way of promoting inter-school competition".

Slovenia

- Eurydice-Slovenia (2009, p. 6): "National assessment provides pupils and their parents with additional information on their level of knowledge, making it comparable with achievements of their peers from other schools and with the national average."

Sweden

- Eurydice (2009, p. 55): "Only a few countries arrange for central government publication of results for each school, or recommend local publication. Such information is published …by the National Agency for Education in Sweden."

Trinidad and Tobago

- Trinidad and Tobago MOE (http://www.moe.gov.tt/student_exam_ntest.html): One goal of the National Test is to "compare students' performance by school and educational districts." To aid this goal "the Ministry of Education publishes an annual report on students' performance on the National Test"

Turkey

- UNESCO (2011): Results from the Education Situation Assessment, which has been in place since 1993, "are used to compare regions, schools and different types of schools in respect to student achievement and to conduct comparisons between new and old programmes."

United Kingdom (England)

- Eurydice-United Kingdom (2009, p. 13): "Schools are required to publish the latest available information showing school and national assessment results in their prospectus." Schools are expected "to publish their targets alongside performance information in their annual prospectus."

United Kingdom (Scotland)

- Eurydice (2009, p. 56-57): "In the United Kingdom (Scotland), the government does not publish school league tables based on results obtained in the certificate examinations held at the end of lower secondary education. However, the results for each school are available on the government website. The press are allowed to use these data if they wish to produce their own school league tables."

Formal Sanction and/or Reward Categorization Rationale

Provided here are key representative references for Formal Sanction and/or Reward country categorization. For other resources used in categorization see appendix C.

Chile

- Manzi et al. (2008): The National System of School Performance Evaluation, in place since 1996, provides monetary incentives to schools that show outstanding performance in several dimensions, including the national SIMCE scores. "Ninety percent of the incentive given to a school is divided among all teachers in the school. Each teacher in turn receives an incentive that typically is half of their monthly salary or between 5% and 7% of their annual salary" (p. 12) Both SIMCE scores and the overall school evaluation scores (SNED) are available to the public.

China (Hong Kong, Macao, and Shanghai)

- Schmidt, Houang & Shakrani (2009, p. 37): "School-level assessment results are published in what are known as 'league ranking tables' for the public to see, with the purpose of increasing competition and school performance. School enrollment quotas are based partially on students' exam scores, so the better the students perform within a school, the more students (and funding) a school can receive".

Hungary

- Eurydice: Hungary (2009): Results from the National Assessment of Basic Competences (NABC) are used to "inform local, regional and national policymakers and the school's clients (parents, students) about school effectiveness" and results are disseminated "to the general public through the schools and other printed and electronic media" (p. 9). "Regulations force schools to continuously evaluate their scores [from NABC]. On one hand, the Public Education Act has an amendment that makes it obligatory for low achieving schools to make provisions; on the other hand the Act pronounces that among other things the annual quality control documents must be based on assessment data" (p. 14).

Latvia

- Eurydice (2009, p. 54): "The results of national tests, aggregated for each school, are considered by the central education authorities in the evaluation or auditing of schools."
- Email Correspondence from Andris Kangro, University of Latvia to Guillermo Montt, OECD: Results are aggregated at the school level and "are used as one of indicators in the process of school accreditation"

Mexico

- OECD – Mexico (2012): The ENLACE exam is administered annually to all students in the third to ninth grade with school level results published and used for teacher evaluation, "for instance, it has a weight of 50% in both the Universal Evaluation System and the National Teacher Career Programme" (p. 40).

Portugal

- Eurydice (2009, p. 54): "The results of national tests, aggregated for each school, are considered by the central education authorities in the evaluation of the auditing of

schools." "Schools with weak results in standardized tests at the ISCED 1 are required to prepare a set of corrective measures and specify their timing. They also propose extra support for underperforming children."

United States

- U.S. Department of Education: Increasing consequences if schools fail to meet adequate yearly progress (AYP) on student achievement scores towards a goal of 100% proficiency in 2014. After four years of failing to reach AYP "the district must implement certain *corrective actions* to improve the school, such as replacing certain staff or fully implementing a new curriculum, while continuing to offer public school choice and supplemental educational services for low-income students." Restructuring of the school occurs after five years.

APPENDIX C: Equations for Stage 1 Models

Model 1: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + ν_{0k}

Model 2: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + δ_{02} (Econ Development) + δ_{03} (Ed Expend) + δ_{04} (Market Competition) + δ_{05} (Market Competition*NTP) + ν_{0k}

Model 3: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + δ_{02} (Econ Development) + δ_{03} (Ed Expend) + δ_{04} (Market Competition) + γ_{1k} (School Type) + γ_{2k} (School Location) + u_{jk} + ν_{0k}

Model 4: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + δ_{02} (Econ Development) + δ_{03} (Ed Expend) + δ_{04} (Market Competition) + γ_{1k} (School Type) + γ_{2k} (School Location) + γ_{3k} (School Mean SES) + u_{jk} + ν_{0k}

Model 5: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + δ_{02} (Econ Development) + δ_{03} (Ed Expend) + δ_{04} (Market Competition) + γ_{1k} (School Type) + γ_{2k} (School Location) + γ_{3k} (School Mean Achievement) + u_{jk} + ν_{0k}

Model 6: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + δ_{02} (Econ Development) + δ_{03} (Ed Expend) + δ_{04} (Market Competition) + γ_{1k} (School Type) + γ_{2k} (School Location) + γ_{3k} (School Mean SES) + γ_{31} (School Mean SES*NTP) + u_{jk} + ν_{0k}

Model 7: School Practice$_{jk}$ = δ_{00} + δ_{01} (NTP) + δ_{02} (Econ Development) + δ_{03} (Ed Expend) + δ_{04} (Market Competition) + γ_{1k} (School Type) + γ_{2k} (School Location) + γ_{3k} (School Mean Achievement) + γ_{31} (School Mean Achievement*NTP) + u_{jk} + ν_{0k}

APPENDIX D: Odds Ratio of National Testing Policy on School Accountability Practices, Controlling for School Mean Math Achievement

Fixed Effects	sc_monitor	sc_comp	p_sc_comp	posted	pr_eval	t_eval
no_ntp	.238* (.139)	.451 (.270)	1.030 (.591)	.357 (.247)	.635 (.370)	.468 (.312)
evaluative	1.825 (.639)	2.350* (.840)	2.091* (.716)	3.095** (1.275)	3.028** (1.053)	2.355* (.937)
form_sanct_rew	2.331 (1.019)	2.440 (1.137)	1.458 (.652)	3.478* (1.868)	1.365 (.619)	1.322 (.686)
n_develop_c	.999* (.001)	.999 (.001)	.999 (.001)	.999 (.001)	.999** (.001)	.999*** (.001)
ed_exp	.679** (.083)	.739* (.092)	.819 (.098)	.672** (.097)	.737* (.090)	.594*** (.083)
n_comp	.975** (.008)	.962*** (.009)	.964*** (.008)	.966** (.010)	.981* (.008)	.978* (.010)
private	.970* (.014)	1.098*** (.013)	.953** (.013)	.700*** (.009)	.868*** (.011)	1.723*** (.021)
sc_loc_town	1.295*** (.016)	1.066*** (.010)	.904*** (.010)	1.210*** (.013)	1.178*** (.012)	1.044** (.010)
sc_loc_city	1.261*** (.017)	.936*** (.010)	.859*** (.010)	1.093*** (.012)	1.214*** (.013)	1.072*** (.012)
sc_loc_largecity	1.263*** (.023)	.979 (.014)	.960** (.015)	1.010 (.015)	1.376*** (.020)	1.168*** (.017)
sc_math_c	1.001*** (.001)	1.001*** (.001)	.999*** (.001)	1.004*** (.001)	.998*** (.001)	.999*** (.001)
Intercept (_cons)	50.389 (32.452)	5.148 (3.379)	.898 (.565)	2.662 (2.015)	2.022 (1.292)	14.336 (10.487)
Random Effects						
Residual (sd_cons)	1.073 (.103)	1.095 (.105)	1.050 (.100)	1.264 (.121)	1.067 (.102)	1.221 (.117)
Model Fit Statistics						
Deviance	297268.3	440751.0	367268.6	405339.4	410664.2	395589.0

Note: Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX E: Odds Ratio of National Testing Policy on School Employing Selective Admission or Transfer Policy (models 4 & 5)

Selective Admission Policy	Model 4	Model 5	**Selective Transfer Policy**	Model 4	Model 5
no_ntp	.657 (.528)	.671 (.523)		1.908 (1.500)	1.937 (1.536)
evaluative	.744 (.356)	.734 (.342)		.568 (.266)	.567 (.268)
form_sanc_rew	1.980 (1.238)	1.456 (.884)		1.851 (1.132)	1.331 (.821)
n_develop_c	.999 (.001)	.999 (.001)		.999** (.001)	.999** (.001)
ed_exp	.483*** (.081)	.498*** (.081)		.679* (.111)	.710* (.118)
n_comp	.985 (.012)	.984** (.011)		1.024* (.012)	1.023 (.012)
private	1.749*** (.024)	1.941*** (.026)		1.883*** (.027)	2.151*** (.030)
sc_loc_town	1.442*** (.015)	1.564*** (.016)		1.344*** (.016)	1.469*** (.017)
sc_loc_city	1.622*** (.019)	1.889*** (.021)		1.161*** (.016)	1.377*** (.018)
sc_loc_largecity	1.558*** (.024)	1.813*** (.027)		1.076*** (.019)	1.299*** (.022)
sc_ses	1.837*** (.015)			2.006*** (.017)	
sc_math_c		1.006*** (.001)			1.007*** (.001)
Intercept (_cons)	49.818 (43.898)	36.261 (30.997)		1.410 (1.216)	.916 (.797)
Random Effects					
Residual (sd_cons)	1.471 (.141)	1.427 (.137)		1.439 (.138)	1.451 (.139)
Model Fit Statistics					
Deviance	399510.9	398612.6		318796.6	317712.1

Note: Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX F: Relationship between National Testing Policy and Availability of Extra-Curricular Activities and School Participating in More than 2 Standardized Tests Annually (models 4 & 5)

Extra-Curricular Activities	Model 4	Model 5	**Standardized Tests (Odds Ratio)**	Model 4	Model 5
no_ntp	-.778 (.959)	-.624 (1.001)		1.529 (1.294)	1.530 (1.271)
evaluative	.769 (.483)	.769 (.504)		1.700 (.856)	1.706 (.846)
form_sanc_rew	1.521** (.576)	1.004 (.601)		2.191 (1.443)	2.327 (1.505)
n_develop_c	1.000 (.001)	1.000 (.001)		.999 (.001)	.999 (.001)
ed_exp	-.893*** (.158)	-.800*** (.165)		.833 (.147)	.838 (.145)
n_comp	-.032** (.012)	-.035** (.012)		.993 (.012)	.993 (.012)
private	.102*** (.014)	.123*** (.013)		1.491*** (.021)	1.531*** (.021)
sc_loc_town	.197*** (.012)	.349*** (.011)		.987 (.011)	.996 (.011)
sc_loc_city	.252*** (.013)	.515*** (.012)		.963** (.012)	.978 (.011)
sc_loc_largecity	.043* (.017)	.328*** (.016)		1.027 (.017)	1.046** (.016)
sc_ses	1.063*** (.008)			.921*** (.007)	
sc_math_c		.010*** (.001)			.998*** (.001)
Intercept (_cons)	11.871 (.811)	11.140 (.846)		.492 (.456)	.482 (.439)
Random Effects					
Residual (sd_cons)	1.293 (.133)	1.350 (.139)		1.548 (.156)	1.520 (.154)
Model Fit Statistics					
Deviance	1359576.8	1361763.4		377588.0	377139.6

Note: **Extra-Curricular Activity**: Unstandardized regression coefficients provided. **Standardized Tests**: Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX G: Estimating Student Grade Repetition from NTP and School Practices

Grade Repetition	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	0.574	0.633	0.701	0.638	0.611	0.624	0.638	0.634	0.648	0.575	0.635	0.658	0.689	0.698
	(0.374)	(0.340)	(0.333)	(0.428)	(0.345)	(0.337)	(0.327)	(0.336)	(0.351)	(0.350)	(0.351)	(0.342)	(0.435)	(0.361)
Form.	1.481	1.504	1.482	1.435	1.247	1.386	1.546	1.508	1.445	1.851	1.479	1.430	2.343*	1.619
San/Rewards	(0.379)	(0.362)	(0.378)	(0.512)	(0.376)	(0.380)	(0.350)	(0.359)	(0.361)	(0.363)	(0.372)	(0.371)	(0.360)	(0.416)
*Eval*Inter*			0.971	0.992	1.085	1.055	1.059	1.009	0.967	1.207	0.978	0.877	0.994	0.988
			(0.028)	(0.112)	(0.060)	(0.062)	(0.061)	(0.044)	(0.091)	(0.138)	(0.177)	(0.136)	(0.030)	(0.038)
*Form.*Inter*			1.014	1.059	1.488***	1.491***	1.002	0.988	1.084	0.782	1.246	1.111	0.958*	1.059
			(0.019)	(0.122)	(0.088)	(0.092)	(0.058)	(0.085)	(0.082)	(0.146)	(0.153)	(0.149)	(0.020)	(0.035)
compliance			1.037**											1.017
			(0.011)											(0.017)
sc_monitor				0.933										0.999
				(0.068)										(0.048)
sc_comp					0.887**									0.993
					(0.037)									(0.053)
p_sc_comp						1.050								1.249***
						(0.034)								(0.054)
posted							0.857**							0.888***
							(0.046)							(0.032)
pr_eval								0.939						0.973
								(0.038)						(0.040)
t_eval									0.954					1.049
									(0.055)					(0.031)
admin										0.676**				0.739***
										(0.109)				(0.072)
transfer											0.681***			0.777**
											(0.085)			(0.074)
S_tests_b												1.127		1.138*
												(0.129)		(0.055)
ex_act													0.937***	0.930***
													(0.018)	(0.014)
female		0.702***	0.702***	0.702***	0.703***	0.703***	0.701***	0.702***	0.702***	0.701***	0.702***	0.702***	0.703***	0.701***
		(0.024)	(0.024)	(0.014)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)
immig		1.237	1.237	1.238	1.237	1.237	1.238	1.237	1.237	1.243	1.237	1.238	1.242	1.245
		(0.151)	(0.152)	(0.151)	(0.152)	(0.152)	(0.152)	(0.151)	(0.151)	(0.155)	(0.151)	(0.151)	(0.158)	(0.163)
lang_other		1.217**	1.217**	1.216***	1.216**	1.215**	1.217**	1.218**	1.216**	1.217**	1.218**	1.215**	1.213**	1.213**
		(0.059)	(0.060)	(0.029)	(0.060)	(0.059)	(0.060)	(0.059)	(0.059)	(0.060)	(0.059)	(0.060)	(0.060)	(0.062)
ses		0.750***	0.750***	0.750***	0.750***	0.751***	0.749***	0.749***	0.750***	0.752***	0.751***	0.749***	0.755***	0.758***
		(0.048)	(0.048)	(0.048)	(0.048)	(0.049)	(0.048)	(0.048)	(0.048)	(0.049)	(0.048)	(0.048)	(0.050)	(0.051)
Intercept	0.145	0.125	0.110	0.132	0.132	0.124	0.129	0.128	0.128	0.152	0.137	0.121	0.193	0.218
	(0.271)	(0.247)	(0.256)	(0.274)	(0.248)	(0.246)	(0.239)	(0.240)	(0.240)	(0.252)	(0.229)	(0.244)	(0.268)	(0.289)
Deviance	73346	63608	63596	63604	63584	63558	63598	63604	63604	63462	63554	63592	63428	63270

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX G: Odds Ratios of National Testing Policy and School Practices on Student's Perception that School Prepares them for Future

CL: Prep for Future	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	1.176	1.146	1.061	1.253	1.206	1.151	1.149	1.186	1.207	1.114	1.157	1.151	1.144	1.001
	(0.156)	(0.163)	(0.167)	(0.173)	(0.160)	(0.161)	(0.163)	(0.162)	(0.161)	(0.166)	(0.160)	(0.159)	(0.177)	(0.183)
Form. San/Rewards	1.054	1.078	1.013	1.114	1.105	1.109	1.068	1.090	1.109	0.960	1.079	1.092	0.767	0.915
	(0.255)	(0.249)	(0.245)	(0.251)	(0.249)	(0.250)	(0.245)	(0.245)	(0.243)	(0.268)	(0.246)	(0.248)	(0.261)	(0.264)
*Eval*Inter*			1.026	0.900*	0.916*	0.984	0.973	0.925*	0.917**	1.071	0.974	0.984	1.000	1.048*
			(0.015)	(0.047)	(0.036)	(0.034)	(0.029)	(0.033)	(0.032)	(0.043)	(0.051)	(0.044)	(0.010)	(0.023)
*Form.*Inter*			1.021*	0.962	0.953	0.888**	1.010	0.977	0.953	1.175*	0.993	0.963	1.040***	1.045*
			(0.010)	(0.038)	(0.030)	(0.042)	(0.025)	(0.051)	(0.052)	(0.071)	(0.050)	(0.039)	(0.008)	(0.020)
compliance			0.979*											0.965**
			(0.009)											(0.012)
sc_monitor				1.030										0.952*
				(0.031)										(0.024)
sc_comp					1.030									0.960*
					(0.024)									(0.017)
p_sc_comp						1.029								0.960
						(0.023)								(0.029)
posted							1.066**							1.029
							(0.022)							(0.016)
pr_eval								1.050*						0.994
								(0.020)						(0.020)
t_eval									1.042*					0.967*
									(0.021)					(0.014)
admin										1.028				1.082*
										(0.025)				(0.031)
transfer											1.024			0.998
											(0.033)			(0.023)
S_tests_b												1.006		0.987
												(0.029)		(0.018)
ex_act													1.012*	1.020**
													(0.006)	(0.006)
female		1.308***	1.308***	1.308***	1.308***	1.308***	1.308***	1.308***	1.308***	1.307***	1.308***	1.308***	1.305***	1.304***
		(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.024)
immig		0.981	0.981	0.982	0.981	0.981	0.981	0.982	0.982	0.978	0.981	0.981	0.981	0.978
		(0.043)	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)	(0.042)	(0.042)	(0.043)	(0.042)	(0.042)	(0.043)	(0.043)
lang_other		0.876**	0.876**	0.877**	0.876**	0.877**	0.877**	0.875**	0.876**	0.877**	0.876**	0.876**	0.878**	0.879**
		(0.043)	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)	(0.042)	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)
ses		1.100***	1.100***	1.100***	1.100***	1.100***	1.100***	1.100***	1.100***	1.097***	1.100***	1.100***	1.094***	1.092***
		(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)
Intercept	2.638	2.456	2.618	2.395	2.418	2.435	2.407	2.406	2.396	2.412	2.433	2.452	2.254	2.456
	(0.115)	(0.121)	(0.127)	(0.125)	(0.121)	(0.121)	(0.122)	(0.119)	(0.120)	(0.122)	(0.120)	(0.120)	(0.127)	(0.149)
Deviance	282676	268926	268916	268920	268918	268916	268908	268920	268920	268872	268926	268926	268802	268798

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX G: Odds Ratios of National Testing Policy and School Practices on Student's Perception that School is not a Waste

CL: Not a Waste	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	0.871	0.798	0.721*	0.848	0.803	0.793	0.793	0.784	0.808	0.812	0.813	0.784	0.889	0.813
	(0.115)	(0.134)	(0.143)	(0.143)	(0.135)	(0.130)	(0.143)	(0.143)	(0.132)	(0.148)	(0.134)	(0.132)	(0.149)	(0.156)
Form. San/Rewards	1.131	1.123	1.049	1.256	1.161	1.145	1.126	1.116	1.142	1.034	1.129	1.141	0.979	1.122
	(0.146)	(0.163)	(0.174)	(0.174)	(0.174)	(0.166)	(0.161)	(0.161)	(0.158)	(0.175)	(0.172)	(0.165)	(0.176)	(0.207)
*Eval*Inter*			1.032	0.927	0.982	1.016	0.982	1.033	0.976	1.017	1.016	1.054	0.985	1.000
			(0.018)	(0.062)	(0.045)	(0.049)	(0.046)	(0.050)	(0.045)	(0.053)	(0.074)	(0.054)	(0.010)	(0.026)
*Form.*Inter*			1.022	0.875*	0.935	0.920	0.969	1.017	0.970	1.102*	0.959	0.955	1.012	0.978
			(0.017)	(0.064)	(0.053)	(0.053)	(0.041)	(0.047)	(0.054)	(0.038)	(0.054)	(0.053)	(0.010)	(0.028)
compliance			0.977											0.997
			(0.013)											(0.017)
sc_monitor				1.101*										1.028
				(0.039)										(0.030)
sc_comp					1.035									0.996
					(0.033)									(0.021)
p_sc_comp						0.999								0.966
						(0.037)								(0.023)
posted							1.094**							1.066**
							(0.029)							(0.023)
pr_eval								1.020						1.010
								(0.037)						(0.023)
t_eval									1.042					0.986
									(0.029)					(0.025)
admin										1.182***				1.174***
										(0.029)				(0.025)
transfer											1.134**			1.089**
											(0.039)			(0.026)
S_tests_b												0.999		0.998
												(0.038)		(0.023)
ex_act													1.034***	1.027***
													(0.007)	(0.005)
female		2.063***	2.064***	2.064***	2.063***	2.063***	2.062***	2.064***	2.064***	2.061***	2.061***	2.063***	2.053***	2.050***
		(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)
immig		1.260**	1.260**	1.260**	1.260**	1.260**	1.259**	1.260**	1.261**	1.247**	1.260**	1.261**	1.258**	1.247**
		(0.079)	(0.079)	(0.079)	(0.079)	(0.079)	(0.080)	(0.079)	(0.079)	(0.076)	(0.079)	(0.080)	(0.080)	(0.079)
lang_other		0.892	0.891	0.892	0.892	0.892	0.894	0.891	0.892	0.894	0.889	0.892	0.897	0.896
		(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.060)	(0.061)	(0.061)	(0.060)
Ses		1.214***	1.214***	1.214***	1.214***	1.214***	1.214***	1.215***	1.214***	1.206***	1.210***	1.214***	1.204***	1.195***
		(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
Intercept	12.230	9.850	10.592	9.100	9.682	9.852	9.584	9.779	9.633	8.874	9.400	9.854	7.700	6.992
	(0.067)	(0.084)	(0.094)	(0.079)	(0.084)	(0.085)	(0.085)	(0.086)	(0.082)	(0.089)	(0.085)	(0.084)	(0.097)	(0.122)
Deviance	30134	20162	20158	20156	20162	20162	20152	20160	20162	20106	20142	20160	20112	20048

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX G: Odds Ratios of National Testing Policy and School Practices on Student's Perception that their Teachers are Interested in their Well-being

CL: Teacher Interested	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	1.156	1.154	1.028	1.247	1.197	1.142	1.127	1.196	1.231	1.079	1.123	1.160	1.017	0.945
	(0.145)	(0.184)	(0.177)	(0.189)	(0.187)	(0.186)	(0.184)	(0.186)	(0.188)	(0.179)	(0.180)	(0.182)	(0.192)	(0.176)
Form.	1.253	1.246	1.170	1.322	1.283	1.240	1.206	1.254	1.250	1.224	1.207	1.251	1.083	1.048
San/Rewards	(0.205)	(0.209)	(0.202)	(0.219)	(0.216)	(0.209)	(0.204)	(0.210)	(0.203)	(0.207)	(0.208)	(0.208)	(0.221)	(0.206)
*Eval*Inter*			1.037**	0.910*	0.932^	1.032	1.072**	0.923**	0.892***	1.125*	1.068	0.978	1.017*	1.066***
			(0.011)	(0.038)	(0.036)	(0.032)	(0.025)	(0.025)	(0.029)	(0.046)	(0.044)	(0.036)	(0.007)	(0.015)
*Form.*Inter*			1.021	0.931	0.943*	1.020	1.104**	0.983	0.993	1.032	1.089*	0.982	1.018**	1.060***
			(0.012)	(0.040)	(0.030)	(0.033)	(0.033)	(0.031)	(0.032)	(0.035)	(0.043)	(0.043)	(0.006)	(0.015)
compliance			0.989											0.970**
			(0.008)											(0.009)
sc_monitor				1.060*										0.982
				(0.026)										(0.020)
sc_comp					1.058*									0.986
					(0.026)									(0.017)
p_sc_comp						0.994								0.975
						(0.027)								(0.016)
posted							0.955**							0.977
							(0.017)							(0.017)
pr_eval								1.021						0.981
								(0.020)						(0.016)
t_eval									1.033*					0.985
									(0.015)					(0.016)
admin										0.938*				0.993
										(0.026)				(0.016)
transfer											0.923**			0.961
											(0.030)			(0.022)
S_tests_b												1.060*		1.057**
												(0.026)		(0.016)
ex_act													0.989*	0.997
													(0.005)	(0.003)
female		1.225***	1.225***	1.226***	1.225***	1.225***	1.225***	1.225***	1.225***	1.226***	1.225***	1.225***	1.226***	1.226***
		(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)
immig		1.047	1.047	1.047	1.046	1.047	1.047	1.047	1.047	1.047	1.047	1.047	1.047	1.048
		(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
lang_other		0.955	0.955	0.955	0.955	0.955	0.955	0.954	0.954	0.954	0.955	0.954	0.954	0.954
		(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
ses		1.032	1.032	1.032	1.032	1.033	1.033	1.032	1.032	1.034	1.034	1.032	1.033	1.034
		(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Intercept	2.212	2.037	2.111	1.942	1.979	2.041	2.063	2.021	2.001	2.121	2.099	2.002	2.213	2.417
	(0.145)	(0.144)	(0.138)	(0.147)	(0.147)	(0.147)	(0.146)	(0.144)	(0.143)	(0.141)	(0.141)	(0.145)	(0.152)	(0.143)
Deviance	290440	277654	277644	277646	277646	277654	277646	277648	277640	277642	277640	277642	277642	277616

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

APPENDIX G: Odds Ratios of National Testing Policy and School Practices on Student's Perception that their Teachers treat them Fairly

CL: Treat Fairly	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evaluative	1.046	1.041	0.910	1.145	1.105	1.036	1.026	1.088	1.082	1.017	1.022	1.069	0.935	0.872
	(0.115)	(0.119)	(0.120)	(0.128)	(0.121)	(0.119)	(0.116)	(0.122)	(0.120)	(0.120)	(0.117)	(0.117)	(0.129)	(0.1128)
Form. San/Rewards	0.844	0.846	0.785	0.842	0.891	0.869	0.832	0.873	0.826	0.803	0.844	0.860	0.686*	0.723
	(0.141)	(0.144)	(0.157)	(0.143)	(0.149)	(0.146)	(0.139)	(0.144)	(0.146)	(0.160)	(0.146)	(0.144)	(0.146)	(0.166)
*Eval*Inter*			1.043**	0.892*	0.894***	1.010	1.033	0.896**	0.935	1.060	1.074	0.936	1.014	1.060**
			(0.013)	(0.048)	(0.031)	(0.035)	(0.035)	(0.040)	(0.043)	(0.051)	(0.050)	(0.044)	(0.007)	(0.022)
*Form.*Inter*			1.024	1.001	0.903*	0.890	1.044	0.919*	1.045	1.075	1.011	0.938*	1.026**	1.047
			(0.016)	(0.038)	(0.041)	(0.060)	(0.035)	(0.033)	(0.041)	(0.053)	(0.041)	(0.032)	(0.007)	(0.024)
compliance			0.977**											0.965**
			(0.007)											(0.011)
sc_monitor				1.070*										0.994
				(0.031)										(0.024)
sc_comp					1.083**									0.980
					(0.024)									(0.018)
p_sc_comp						1.034								0.969
						(0.022)								(0.029)
posted							0.999							0.991
							(0.027)							(0.017)
pr_eval								1.075**						1.001
								(0.027)						(0.021)
t_eval									1.001					0.957
									(0.029)					(0.027)
admin										1.026				1.064**
										(0.038)				(0.021)
transfer											0.976			0.983
											(0.033)			(0.019)
S_tests_b												1.078**		1.031
												(0.027)		(0.017)
ex_act													1.002	1.011**
													(0.006)	(0.004)
female		1.304***	1.305***	1.305***	1.305***	1.304***	1.304***	1.305***	1.304***	1.304***	1.304***	1.305***	1.303***	1.303***
		(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.03)
immig		1.025	1.025	1.025	1.025	1.025	1.025	1.026	1.026	1.022	1.025	1.025	1.025	1.022
		(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.041)	(0.042)	(0.042)	(0.043)	(0.042)
lang_other		0.915*	0.915*	0.916*	0.915*	0.915*	0.915*	0.913*	0.915*	0.915*	0.915*	0.914*	0.915*	0.916*
		(0.041)	(0.041)	(0.042)	(0.042)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
ses		1.029*	1.029*	1.029*	1.029*	1.029	1.029*	1.029*	1.029*	1.027	1.029*	1.029*	1.026	1.024
		(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Intercept	4.232	3.843	4.131	3.631	3.691	3.808	3.844	3.740	3.841	3.779	3.877	3.763	3.790	4.006
	(0.086)	(0.090)	(0.096)	(0.096)	(0.091)	(0.090)	(0.087)	(0.087)	(0.085)	(0.089)	(0.088)	(0.091)	(0.085)	(0.109)
Deviance	239566	227204	227192	227194	227190	227192	227200	227194	227196	227184	227200	227196	227164	227150

Note: **Bold** identifies variable included in interaction. Population-average model used in HLM6 software. Odds ratios provided. Robust standard errors in parentheses. *p<.05 **p<.01 ***p<.001

VITA

William C. Smith

EDUCATION

2014 The Pennsylvania State University
Dual PhD in Education Theory and Policy and Comparative International Education Minor: Sociology Concentration: Quantitative Methods

2011 Josef Korbel School of International Studies: University of Denver
MA in International Development Concentrations: Political Theory and Education

2003 Western Oregon University
MA in Teaching Concentration: Secondary Social Studies

2001 Portland State University
BS in Sociology with High Honors Minor: Psychology

SELECT GRANTS

2014 Program for the International Assessment of Adult Competencies (PIACC) Research Grant ($8,000) – American Institutes of Research (AIR)
Project Title: A Comparative Study of Immigrant and Native Employees in the United States and Canada

2013 Thomas J. Alexander Fellowship ($35,000) – OECD

SELECT PUBLICATIONS

Smith, W. (2014). The global transformation toward testing for accountability. *Education Policy and Analysis Archives*.

Smith, W. et al. (2014). A meta-analysis of the effects of education on chronic disease: The causal dynamics of the Population Education Transition Curve. *Social Science and Medicine.* http://dx.doi.org/10.1016/j.socscimed.2014.10.027

Smith, W. (2014). Estimating unbiased treatment effects in education using a regression discontinuity design. *Practical Assessment, Research & Evaluation*, 19(9).

Smith, W. & Rowland, J. (2014). Parent trigger laws and the promise of parental voice. *Journal of School Choice*, 8(1), 94-112.

Smith, W. (2013). Framing the debate over teacher unions. *Mid-Atlantic Education Review* 1(1), 17-26.

Smith, W., Salinas, D. & Baker, D. (2012). Multiple effects of education on disease: The intriguing case of HIV/AIDS in Sub-Saharan Africa. In Wiseman, A. & Glover, R. (Eds.), *The Impact of HIV/AIDS on Education Worldwide* (pp. 79-104).

Joshi, D. & Smith, W. (2012). Education and inequality: Implications of the World Bank's Education Strategy 2020. In Wiseman, A. & Collins, C. (Eds.), *Education Strategy in the Developing World: Revising the World Bank's Education Policy* (pp. 173-202).

Halabi, S. Smith, W., Collins, J., Baker, D. & Bedford, J. (2012). A document analysis of HIV/AIDS education interventions in Ghana. *Health Education Journal*.