EXAMINING THE INFLUENCES OF YUTORI EDUCATION IN JAPAN ON OPPORTUNITY TO LEARN (OTL) AND STUDENT ACHIEVEMENT ON THE TIMSS: A MULTIPLE COHORT ANALYSIS

by

Meiko Lin

Dissertation Committee:

Professor Madhabi Chatterji, Sponsor
Professor Oren Pizmony-Levy

Approved by the Committee on the Degree of Doctor of Education

Date May 16, 2018

Submitted in partial fulfillment of the
Requirements for the Degree of Doctor of Education in
Teachers College, Columbia University

2018

ABSTRACT

EXAMINING THE INFLUENCES OF YUTORI EDUCATION IN JAPAN ON OPPORTUNITY TO LEARN (OTL) AND STUDENT ACHIEVEMENT ON THE TIMSS: A MULTIPLE COHORT ANALYSIS

Meiko Lin

The purpose of this study was to explore the effects of *yutori* reforms on Opportunity to Learn (OTL), as defined by Stevens' (1993, 1996) multidimensional framework, and to examine how the changes in OTL may have subsequently affected Japanese 8th graders' mathematics achievement as measured by the Trends in Mathematics and Science Study (TIMSS). This dissertation was a mixed-methods, multi-cohort study combining analyses of archival documents and interview-based data with analyses of quantitative TIMSS data on OTL and student achievement in mathematics in selected years. The study used three waves of TIMSS data (1999, 2003, and 2007) to examine the effects of *yutori* reforms on OTL levels at the classroom level over time, and their corresponding influence on student achievement levels on the TIMSS assessment with Hierarchical Linear Models (HLM).

The three overarching findings of this study were: (a) the *yutori* reforms were not implemented in schools and classes as originally intended by the Ministry of Education, Sports, Culture and Technology in Japan, with ongoing shifts in policies and priorities at the national level; (b) there were significant changes in classroom-level OTL measures, indicating reductions in instructional time dedicated to mathematics but improvements in the quality of instructional delivery were found to occur under the *yutori* reforms; and (c) the instructional time component of OTL was found to be positively associated with students' mathematics achievement under *yutori* reforms, with the most socioeconomically disadvantaged students benefitting more in terms of achievement outcomes than those who were more advantaged.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the tutelage and support of my advisor, Dr. Madhabi Chatterji. Thank you for tirelessly providing me with guidance, insight, and feedback throughout the dissertation process. You challenged me with your high expectations for my work, and I am fortunate for all the ways you pushed me to do my best.

I also must thank my committee members, Dr. Oren Pizmony-Levy, Dr. Lyle Yorks, and Dr. Stuart Weinberg, for providing substantive feedback and challenging my thinking. Your input was invaluable in further strengthening the manuscript.

To Dena and Sukhminder, thank you so much for being my friends and thought-partners. You were always generous with your time and provided me with the support that I need. I am grateful to have friends like you. To Marisol, I feel so privileged and blessed to know you. You have always supported me and cheered me on. I will always cherish our long study sessions and our friendship. To Arushi, Jong Ming, and Noriko, I am grateful to have you as dear, dear friends in my life. You have always supported me and cheered me on.

Lastly, I would like to acknowledge my family, and my cousins. I could not have finished this dissertation without your unconditional support and encouragement.

M. L.

TABLE OF CONTENTS

LIST OF TABLES

Table

Table

LIST OF FIGURES

Figure

Chapter I

INTRODUCTION

Yutori Reforms and Opportunity to Learn

Yutori Reforms to Liberalize Japanese Education

In April 2002, the Japanese Education Ministry (Ministry of Education, Sports, Culture and Technology or MEXT) introduced a nation-wide education reform program for primary, junior high, and high school levels. This reform initiative was called "*yutori-kyoiku*" which denotes a "*relaxed education,*" emphasizing the principle that students need a liberal, flexible, and comfortable school life to develop their individuality (MEXT, 2000). The *Yutori-kyoiku* policy was proposed in response to concerns about the ill effects of testing and academic pressure that had been building for decades in schools. Japanese students may have been scoring well on standardized examinations, but their interest in learning was diminishing (Kariya & Rappleye, 2010).

The intensified academic pressure resulted in some students being left behind. In particular, the entrance examinations for high school and college—operated by agencies responsible for the so-called "*examination hell"* (Tsuneyoshi, 2004, p. 368)—were accused of putting excessive pressure on students and destroying their interest in learning (Bjork & Tsuneyoshi, 2005; Rohlen, 1983). Many children had no future aspirations and little desire to progress in school, as evidenced by the increase in the number of young people who were not engaged in any form of employment or education (The Organisation for Economic Co-operation and Development [OECD], 2012b). Rather than emphasize the transmission of large amounts of knowledge through rote learning, MEXT wanted to promote child-initiated learning through the *yutori* reforms.

Yutori curricular reforms constituted a 30% cut in instruction in core academic subjects (i.e., Japanese, mathematics, science), the implementation of a 5-day school week instead of a 6-day school week, and the introduction of "*integrated studies*" (IS) classes dedicated to student-centered, experiential learning in elective subjects (MEXT, 2002). The IS classes aimed to cultivate students' ability to discover, ask questions, and develop the ability to learn, think, and decide independently (MEXT, 2002). Schools or teachers were given flexibility to determine the length of IS classes and to select topics that match the interests of their students and the unique characteristics of their communities (Bjork & Tsuneyoshi, 2005). In an attempt to devolve authority over the education system to local levels, MEXT deliberately kept its directives and reform guidelines for *yutori* reforms at a minimum. As one MEXT official explained, "We wanted the teachers to create the curriculum by themselves, producing their original ideas without the influence of the ministry" (Bjork, 2015, p. 28).

Yutori **Reform Policies and Opportunity to Learn**

The *yutori* reforms cut the hours of instruction in traditional subjects in primary, middle, and secondary schools to make time for IS classes. Tables 1 and 2 summarize the changes in dedicated time to curriculum delivery before and after the *yutori* reforms. Between 1992 and 2002, at the elementary school level, time for teaching all subjects (except for ethics and special activities) was cut from 14% to 18%. At middle school level, at least 70 hours per year were expected to be dedicated to IS classes. To make time for this expansion, hours for core subjects (i.e., Japanese, mathematics, science) were reduced by 18% or more. The instructional hours for mathematics could be reduced as much as 25% in a school year (see Table 3). Given that students had less time for learning the core subjects at school, one may argue that Japanese students had less Opportunity to Learn (OTL) the core academic subjects (Cave, 2003).

OTL generally refers to what students have a chance to learn (Floden, 2002). OTL as a construct was originally conceptualized to indicate whether students had sufficient time to learn and received adequate instruction in a particular domain (Carroll, 1963). Over the last 3 decades, definitions of the construct have expanded beyond the time variable, to incorporate content coverage and quality of instruction (Brewer & Stasz, 1996; Floden, 2002; McDonnell, 1995; Porter, 1991; Stevens, 1996). OTL is fundamental to the goals of schooling, given that schools are "organized to provide students with the time and experiences geared towards learning and mastery of specific subject matter" (Cogan & Schmidt, 2015, p. 207).

A vast body of research literature has shown OTL, defined in terms of instructional time, content coverage, and quality of instruction, to be a positive correlate of student achievement (Hattie, 2009; Marzano, 2003; Scheerens & Bosker, 1997; Scheerens, Luyten, Steen, & Luyten-De Thouars, 2007). Given the strong OTL-achievement association, opponents of the *yutori* reforms predicted from the very beginning of the reforms that the "watering down" of the curriculum and reduced school hours would undermine the academic performance of Japanese students (Cave, 2003;

Tsuneyoshi, 2004). These critics believed that MEXT was improving creativity and fostering independent learning at the expense of deterioration of measured knowledge levels of students.

Table 1

Changes in Instructional Time for Elementary School Students Leading up to the 2002 Yutori Reforms

Subject Area	Pre-*yutori* (1992)	During *yutori* (2002)	% Change
Japanese	1601	1377	-14%
Social studies	420	345	-18%
Mathematics	1011	869	-14%
Science	420	350	-17%
Music	418	358	-14%
Fine arts	418	358	-14%
Physical education	627	540	-14%
Home economics	627	540	-14%
Ethics	209	209	0%
Special activities	209	209	0%
Integrated Studies	-	430	-
All Subjects-Total	5785	5367	-7%

Note: Standard class times are 50 minutes for middle schools. The instructional time presented are the total instructional hours from Grade 1 to Grade 6. Parentheses ( ) contain the year in which data were collected. Adapted from Bjork, C. (2016). *High-stakes schooling.* Chicago, IL: The University of Chicago Press.

Table 2

Changes in Instructional Time for Middle School Students Leading up to the 2002 Yutori Reforms

Subject Area	Pre-*yutori* (1992)	During *yutori* (2002)	% Change
Japanese	455	350	-23%
Social studies	350-385	295	-16% ~ -23%
Mathematics	385	315	-18%
Science	315-350	290	-8% ~ -17%
Music	140-175	115	-18% ~ -34%
Fine arts	140-175	115	-18% ~ -34%
Physical education	315-350	270	-14% ~ -23%
Technology/home economics	210-245	175	-17% ~ -29%
Foreign language	-	315	-
Moral education	105	105	0%
Special activities	105-210	105	0% ~ -50%
Electives	-	155-280	-
Integrated Studies	-	210-335	-
Total	3150	2940	-7%

Note: Standard class times are 50 minutes for middle schools. The instructional time presented are the total instructional hours from Grade 7 to Grade 9. Parentheses ( ) contain the year in which data were collected. Adapted from Bjork, C. (2016). *High-stakes schooling.* Chicago, IL: The University of Chicago Press.

Table 3

Changes in Instructional Hours for Mathematics Leading up to the 2002 Yutori Reforms

	Pre-*yutori* (Before 2002)	During *yutori* (2002-2003)	% Change
Grade 1	136	114	-16%
Grade 2	175	155	-11%
Grade 3	175	150	-14%
Grade 4-6	175	150	-14%
Grade 7	105	105	0%
Grade 8-9	140	105	-25%

Note: Source: National Institute for Education Research (NIER), 1989; Monbusho, 1990, 2000a; Japan Society of Mathematical Education, 2000. Parentheses ( ) contain the year in which data were collected.

Student Performance Levels Under *Yutori* Reforms

Adoption of these sweeping curricular reform initiatives drew heavy criticism even before they were fully implemented (Tsuneyoshi, 2004). Critics from diverse sectors began to link the *yutori* reforms to the lowering of academic expectations in Japanese students from the end of the 1990s, and even as the new curriculum was being implemented in 2002. According to this view, Japanese students were studying less than before, or at least less than their peers in other industrialized countries (Kariya & Shimizu, 2004; Stevenson & Stigler, 1992). The decreasing hours of study, together with the reduced pressure of the university entrance examinations due to a declining population of 18-year-olds, were equated with a lowering of academic standards (Kariya & Shimizu, 2004; Tsuneyoshi, 2004).

It was within the above context that a sensationalized "achievement crisis" rhetoric emerged among education observers: "If 30% of the curriculum content was reduced and schools hours were reduced as well,…it seems common sense to assume that achievement will suffer" (Tsuneyoshi, 2004, p. 388).

This "standards crisis" debate persisted despite a lack of reliable longitudinal data for assessing the students' scholastic achievement trends through domestic programs (Takayama, 2007, 2008; Tsuneyoshi, 2004). Given the data shortage, international student assessment results generated by the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) were used extensively by anti-reform critics to "confirm Japanese children's achievement crisis" (Takayama, 2007, p. 436).

In December 2004, the release of the PISA 2003 and TIMSS 2003 results sent shockwaves throughout Japan, according to media outlets (Asahi, 2004a, 2004b; Nikkei, 2004; Yomiuri, 2004a, 2004b, 2004c). On international league tables, the average performance of Japanese 15-year-olds dropped from 1st to 6th rank in mathematics and from 8th to 14th in reading on the PISA 2003 tests. Compared to their performance on the PISA 2000, the Japanese students' mean performance dropped from 557 to 543 in mathematics, from 522 to 498 in reading, and from 550 to 548 in science. In particular, the Japanese reading literacy scores on the PISA 2003 test were lower than those on the PISA 2000, and this difference was statistically significant at the 5% error level (OECD, 2004; Takayama, 2008). Though Japanese students performed slightly better than peers in other nations on the TIMSS 2003, there was also a decline in the mean mathematics (from 597 to 565 among 4th graders; from 579 to 570 among 8th graders) and science test scores (from 574 to 543 among 4th graders) within Japan over time. With respect to learning motivation, 29% of the 4th graders and 9% of the 8th graders tested on the TIMSS 2003 tests strongly agreed with the statement "*I enjoy learning mathematics*," which was lower compared to the corresponding international averages of 50% and 29% (Mullis, Martin, Gonzalez, & Chrostowski, 2004, pp. 159-160). Soon after the press release of these findings, the then-minister of education, Nariaki Nakayama, publicly acknowledged a declining scholastic achievement trend in Japan (Takayama, 2008).

The mean mathematics test scores and international ranking of Japanese students dropped between PISA 2000 and PISA 2003. But, the interpretation that the Japanese education system was in crisis or failing was unfounded (Takayama, 2008). There was actually no statistically significant difference between Japan and top-performing Hong

Kong insofar as the mathematical literacy test results were concerned (see OECD, 2004). Moreover, there was no statistically significant change in Japanese 15-year-olds mathematical literacy levels, based on results of PISA 2000 to PISA 2003 (see OECD, 2004). The phenomenon of low learning motivation, observed in the TIMSS 2003, was nothing new for the nation because Japanese students had historically exhibited higher academic performance and lower motivation levels for learning mathematics on surveys, as compared to other nations, since the First International Mathematics Study (FIMS; Takayama, 2008).

Furthermore, on the surveys, most parents indicated concern that their children were not getting enough content and instructional time at the public schools under *yutori* reforms. According to a 2008 survey by MEXT, 67% of parents believed that "school classes alone were not sufficient" and another 14% believed that "school classes alone could not adequately prepare children for school entrance exams" (MEXT, 2008). Such dissatisfaction with schools motivated parents to rely on *juku*, which are private, after-school tutoring institutions, to help their children advance through the educational system and succeed on high school and university entrance examinations.

At the primary school level, *juku* participation increased by 3% to 5%, and at the middle school level by 2% between 2002 and 2007 (see Figure 1). Through Grades 2 to 9, *juku* attendance rose monotonically from Grade 1 to Grade 9 over the span of four surveys (1985, 1993, 2002, 2007). Generally, *juku* participation was growing between 1985 to 1993. *Juku* enrollments have slightly declined across almost all grades in 2002, partly due to the competition from other forms of out-of-school opportunities to receive academic instruction, namely distance learning. Distance education provides study materials to its subscribers, who receive tutoring, complete worksheets, and take practice tests at home via the internet (OECD, 2011). Though distance education was not available until 1999, it nevertheless accounted for 11.7% of supplemental education among elementary school students and 11.8% among middle school students in 2002 (MEXT, 2008). A major reason behind the growing popularity of distance learning programs is their cost, which could be as low as a quarter of the cost of *juku.*

A major concern related to *juku* is the financial burden it places on families. According to OECD (2012), the average expenditure per student in Japan more than doubled between 1985 and 2007, reaching an average of $3,150 U.S. dollars annually, around 7% of per capita income. Such heavy financial burdens imposed by *juku* could further widen educational inequalities between the students from low socioeconomic backgrounds and those from higher socioeconomic backgrounds who can afford the *jukus* (Bjork, 2009; OECD, 2012; Takayama, 2008).

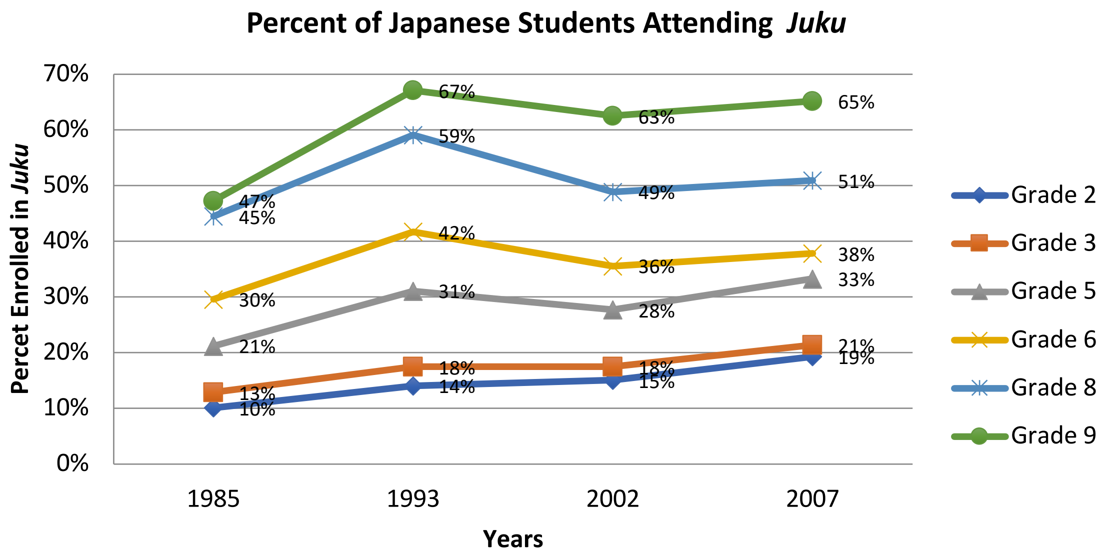


Figure 1. Percent of Japanese students attending *juku*
before and after initiation of *yutori* reforms in 2002
Data from 2008 Student and Extracurricular Activities Survey (MEXT, 2008)

Unresolved Issues in Research on OTL and International Assessments

Given the theoretical implications of *yutori* reform policies on students' OTL, this study was concerned mainly with variations in OTL levels in the context of *yutori* reforms in Japan, and their possible influence on levels of student achievement in multiple, cross-sectional TIMSS cohorts assessed over time.

Between the time *yutori* was first announced (in 1998) and the time *yutori* was officially terminated (in 2011), Japan offered an opportunity for a within-country and in-depth, longitudinal study of *yutori* reforms. Because of a substantial reduction in curricular content and instructional time for core academic subjects (i.e., Japanese, mathematics, science) for all public elementary and middle school students, *yutori* reforms constituted a natural experiment that prompted interesting yet unanswered questions on the possible impact of reforms on OTL levels in schools and classrooms, and their corresponding influence on student achievement levels on the TIMSS, a testing program in which Japan routinely participates. While much of the international assessment literature (Husén, 1967; Kifer & Burstein, 1992; Schmidt & Burstein, 1993; Schmidt & Maier, 2009; Schmidt, McKnight, Houang, & Wang, 2001; Schmidt et al., 2013) has emphasized between-country comparisons of student performance to make interpretations of school effectiveness, deeper within-country analyses are largely overlooked by the research community. This study sought to fill this specific gap in the theoretical and policy literature on International Large-Scale Assessments (ILSA) by looking at Japan as a case over time. The study examined the intersection of relevant constructs at three levels over time (Gonzalez & Miles, 2001; Mullis, Martin, & Foy, 2008; Mullis et al., 2004):

1. *Level 1: The intended curriculum*, as given through perspectives of Japanese government officials at the highest levels of national educational policymaking;
2. *Level 2: The implemented curriculum*, as evidenced through observations of school-based researchers, and as reported by teachers on the TIMSS Teacher Questionnaire; and
3. *Level 3: The attained curriculum*, as evidenced through student performance on TIMSS mathematics achievement tests.

As such, the present study provides a window into the potential relationships of OTL, operationally defined with a multidimensional framework, the facilitators and barriers to *yutori* reform implementation, and changes in student achievement levels as the nation was undergoing sweeping curricular reforms.

Educators and researchers have studied OTL for nearly half a century. Starting in the 1960s, separate OTL research strands emerged around three different aspects of instruction: time given to instruction (Carroll, 1963), content of instruction (Husén, 1967), and quality of instruction (Brophy & Good, 1986). Finally, a multidimensional OTL framework was proposed in the early 1990s (Stevens, 1993).

OTL research on time given to instruction was first introduced through John Carroll's (1963) model of school learning. This strand of research focused on the amount of instructional time needed for students to receive adequate instruction to learn and asks the question: "*How long* will it take this person to learn?" (Floden, 2002, p. 232). Many studies have shown dedicated instructional time to be a strong contributor to student achievement in many academic subjects (Berliner, 1978, 1990; Denham & Lieberman, 1980; Fisher et al., 1981; Harnischfeger & Wiley, 1985; Karweit, 1985; Walberg & Frederick, 1982).

Also in the 1960s, separate from Carroll's work, a second OTL research strand emerged on content of instruction with the First International Mathematics Study (FIMS). FIMS was launched by the International Association for the Evaluation of Educational Achievement (IEA) in 1963 to investigate the outcomes of school systems in various countries (Husén, 1967). FIMS framed OTL as a content-coverage variable without specific regard to instructional time (Floden, 2002; McDonnell, 1995). The FIMS, along with subsequent IEA and other international comparative studies, reported a positive relationship between students' curricular exposure and their achievement (Carnoy, Khavenson, Loyalka, Schmidt, & Zakharov, 2016; Floden, 2002; Husén, 1967; Kifer & Burstein, 1992; Schmidt & Burroughs, 2013; Schmidt & Maier, 2009; Schmidt, Zoido, & Cogan, 2013; Travers & Westbury, 1989).

OTL as quality of instruction is the focus of the third strand of OTL research which emerged as a result of the operationalization of quality of instructional practices such as direct instruction, guided feedback, student think-alouds, and instructional grouping formats (Elliott & Bartlett, n.d.). Past research has found strong and positive effects of instructional strategies on student achievement (Brophy & Good, 1986; Saxe, Gearhart, & Seltzer, 1999; Walberg, 1986).

The three OTL research strands collectively suggest that OTL is an evolving construct from which better understandings may be gleaned on the intricacies of the schooling processes and student achievement (Elliott & Bartlett, n.d.). Progress in OTL research has influenced many researchers to think of, and to operationalize OTL, as a teacher effect in analytic models. At the same time, the confluence of different formulations of OTL became appealing to policymakers in the United States (McDonnell, 1995; Schmidt & Maier, 2009). Researchers and policymakers proposed to broaden the OTL concept to represent school processes and classroom activities so that schools and teachers could have a vision of how to improve (McDonnell, 1995; Stevens, 1996).

Despite the numerous studies on OTL, Stevens (1993) pointed out a critical gap in the OTL research. Namely, the issue was that most OTL studies "looked at one variable at a time" (p. 4). None of the previous studies evaluated effects of OTL on student achievement with OTL as a multidimensional framework (Wang, 1998). Stevens (1993, 1996) went on to offer the first multidimensional conceptual framework of OTL, identifying four OTL elements that have been prevalent in research: *content coverage, content exposure, content emphasis,* and *quality of instructional delivery.* Stevens' framework offered a way to examine the aforementioned three OTL research strands from multidimensional perspectives. Most importantly, Stevens (1993, 1996) treated OTL as a teacher-level variable related to the allocation of adequate instructional time covering a core curriculum, with different cognitive demands and instructional practices that could produce student achievement.

Problem Statement

Purpose

The purpose of this study was to explore the effects of *yutori* reforms on OTL, as defined by Stevens' (1993, 1996) multidimensional framework, and examine how the changes in OTL may have subsequently affected Japanese 8th graders' mathematics achievement as measured by the TIMSS 1999, 2003, and 2007 mathematics assessments. The focus on students' mathematics achievement reflects the enduring emphasis in Japanese national policies on this subject. Given that mathematics is one of the core academic subjects that has experienced the severest cut in instructional hours and that the mathematics achievement of the nation's students is well-documented in ILSAs, the study chose to focus on mathematics.

Specifically, the study presents both classroom-level and multilevel analysis of curricular reforms over time in a single nation, using three cohorts of nationally representative 8th grade student samples. The three cohorts were identified from three consecutive rounds of TIMSS data: (a) pre-*yutori* cohort (i.e., students who studied at Grade 8 and participated in the TIMSS in 1999, and their teachers/administrators); (b) mid-*yutori* cohort (i.e., students who studied at Grade 8 and participated in the TIMSS in 2003, and their teachers/ administrators); and (c) post-*yutori* cohort (i.e., students who studied at Grade 8 and participated in the TIMSS in 2007, and their teachers/ administrators). Of the cohorts, the post-*yutori* cohort was educated entirely under the *yutori* initiative, whereas the mid-*yutori* cohort attended Grades 1-7 before the *yutori* reforms were in full effect.

Research Questions

Six questions were examined, organized by the TIMSS three-tier model of the curriculum: *intended*, *implemented*, and *attained* (Gonzalez & Miles, 2001; Mullis et al., 2008; Mullis et al., 2004).

Intended Curriculum: Research Questions

1. Based on a content analysis of white papers published by MEXT between 1999 and 2011:
    a. What was the *intended curriculum* as given by MEXT?
    b. In what ways did the motivations and intentions of MEXT change, if at all, through the pre-*yutori*, mid-*yutori*, and post-*yutori* time periods?
2. Based on the perceptions of National Research Coordinators (NRC) as reported on the TIMSS surveys in 1999, 2003, and 2007, to what extent did the intended national mathematics curriculum change in terms of content coverage as yutori reforms were implemented between 2003-2007 in Japan?

Implemented Curriculum: Research Questions

3. From two school-based researchers' observations of Japanese junior high schools during the *yutori* reform period:
    a. What roles did teachers and schools play in implementing the *yutori* reform directives?
    b. What support and barriers did schools and teachers face as they attempted to follow *yutori* reform guidelines?
    c. What were other contextual factors affecting the implementation of *yutori* reforms in Japanese schools and classrooms?
4. To what extent is Stevens' multidimensional framework suggesting four interrelated OTL constructs, upheld in the TIMSS 1999, 2003, and 2007 survey data from samples of participating Japanese 8th grade mathematics teachers?

5. Using the validated OTL measures validated against Stevens' framework, to what extent did the *yutori* curricular reforms affect changes in OTL levels over time in 8th grade mathematics classes, as given by the TIMSS teacher survey data from 1999, 2003, and 2007?

Attained Curriculum: **Research Questions**

6. To what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement?
    a. At the classroom level, to what extent did the observed changes in OTL levels over time affect changes in aggregated 8th grade students' mathematics achievement between cohorts (pre-*yutori,* mid-*yutori,* and post-*yutori*)?
    b. Using a multilevel modeling approach, to what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement within cohorts?
    c. Did OTL moderate the relationship between students' socioeconomic backgrounds—a background factor expected to affect *juku* participation levels—and mathematics achievement within cohorts, as measured by the TIMSS student assessments?

Methodological Rationale

To examine the above problems, I employed a mixed-methods multiple cohort study, combining analyses of archival documents and interview-based data with analyses of quantitative TIMSS data on OTL and student achievement in mathematics in selected years. Teacher surveys were used as the main data source for deriving and validating theoretically-supported OTL measures, supplemented with the TIMSS survey data from national education administrators and students.

I used three waves of the TIMSS data (1999, 2003, and 2007) to examine the effects of *yutori* reforms first on OTL levels in classrooms, and then their corresponding influence on student achievement levels on the TIMSS over time. Even though *yutori* reforms were in effect until 2011, I did not include the TIMSS 2011 data due to limited OTL-related items. The TIMSS 2011 Teacher Questionnaire had 19 questions related to topics covered in mathematics class, whereas the previous three waves of the TIMSS had at least 34 topic-coverage questions. Considering the small overlap in *content coverage* items between the TIMSS 2011 data and the previous three rounds of the TIMSS data, I decided to exclude the TIMSS 2011 data from this study.

To address the specific questions that follow, I employed a convergent parallel, mixed-methods design (Creswell & Clark, 2018). A mixed-methods design was selected because such research can provide “a more complete understanding of a research problem than either quantitative or qualitative research alone” (Creswell, 2014, p. 19). The convergent parallel design is best applied when “the researcher intends to bring together the results of the quantitative and qualitative data analysis so they can be compared or combined” (Creswell & Clark, 2018, p. 65). More importantly, the convergent parallel design allows the researcher to validate one set of findings with the other and to gain a more complete understanding of an issue (Creswell & Clark, 2018).

Investigating the *Intended Curriculum* Under *Yutori* Reforms

This study consisted of two qualitative phases followed by two quantitative analytic phases. In the first qualitative phase, I examined and analyzed annual white papers published by MEXT between 1999 and 2011 to map out key components of *yutori* reforms and their implementation timeline, as intended by the Japanese government. I further analyzed descriptive data from the TIMSS NRC Questionnaire to depict the range of topics intended to be covered in 8th grade mathematics curriculum in Japan by year (TIMSS, 1999, 2003, 2007).

Examining the *Implemented Curriculum* Under *Yutori* Reforms

In the second qualitative phase, I conducted two semi-structured interviews with two researchers who have studied Japanese education for more than 10 years and have done observational research in elementary or junior high schools during the *yutori* reform period. The purpose of the interviews was to explore and understand how *yutori* reforms were actually implemented in schools and classrooms.

In the first quantitative phase, I derived OTL construct measures drawing on Stevens' OTL framework from the TIMSS Teacher Questionnaire, and validated the OTL construct measures using an iterative process (Chatterji, 2003; in press). This user-centered, iterative scale design and validation methodology was guided by the intended inferences and uses to be made with construct measures, whether in research or applied contexts. It was recently applied to designing and validating non-cognitive measures for elementary school children by Chatterji and Lin (2018). The relevance of uses in instrument design and validation efforts was highlighted in the latest Standards for Educational and Psychological Testing (see AERA, APA, & NCME, 2014).

The OTL measures were derived as these allowed a measurement of change by year (TIMSS, 1999, 2003, 2007) on all validated OTL dimensions. In the next phase, a quantitative phase, I first described the changes in the range of topics taught in 8th grade mathematics curriculum, using quantitative data collected from the TIMSS Teacher Questionnaire. After that, I investigated the effects of *yutori* reforms on Japanese 8th grade teachers' self-reports of validated OTL measures in mathematics classrooms between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori* cohort).

Evaluating the *Attained Curriculum* Under *Yutori* Reforms

The mathematics achievement of 8th graders as measured on the TIMSS Student Assessment was operationally defined as the *attained curriculum*. In the second quantitative phase, I examined the effects of *yutori* reforms on the OTL and mathematics achievement relationship in three stages: (a) I compared the OTL-mathematics achievement relationship between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori cohort*); (b) I examined the effects of OTL on students' mathematics achievement within cohorts; and (c) I investigated the moderating effects of SES on the relationship between students' mathematics achievement and OTL within cohorts.

Finally, the quantitative and qualitative findings were compared to see in what ways the results converged and diverged. An overview of the research questions aligned to data sources and analytic methods is presented in Table 4, with further details given in Chapter III.

Table 4

Research Questions and Overview of Methods

Curriculum level addressed	Research question	Data sources	Method of analysis
The *intended curriculum*	1. Based on a content analysis of white papers published by MEXT between 1999 and 2011: a. What was the *intended curriculum* as given by MEXT? b. Why did the motivations and intentions of MEXT change through the pre-*yutori*, mid-*yutori*, and post-*yutori* time periods?	White papers published by MEXT between 1999 and 2011	Content analysis (Qualitative)
	2. Based on the perceptions of National Research Coordinators (NRC) as reported on the TIMSS surveys in 1999, 2003, and 2007, to what extent did the intended national mathematics curriculum change in terms of content coverage as yutori reforms were implemented between 2003-2007 in Japan?	TIMSS NRC Questionnaire data from TIMSS 1999, 2003, and 2007	Descriptive statistical analysis on topic coverage (Quantitative)
The *implemented curriculum*	3. From two school-based researchers' observations of Japanese junior high schools during *yutori* reform period: a. What roles did teachers and schools play in implementing *yutori* reform directives? b. What support and barriers did schools and teachers face as they attempted to follow *yutori* reform guidelines? c. What were other contextual factors affecting the implementation of *yutori* reform in Japanese schools and classrooms?	Interview data	Content analysis (Qualitative)
	4. To what extent is Stevens' multidimensional framework suggesting four inter-related OTL constructs, upheld in the TIMSS 1999, 2003, and 2007 survey data from samples of participating Japanese 8th mathematics teachers?	TIMSS Teacher Questionnaire data from TIMSS 1999, 2003, and 2007	Content validation and empirical validation (Quantitative)
	5. Using the validated OTL measures per Stevens' framework, to what extent did the *yutori* curricular reforms affect changes in OTL levels over time in 8th grade mathematics classes, as given by the TIMSS teacher survey data from 1999, 2003, and 2007?	TIMSS Teacher Questionnaire data from TIMSS 1999, 2003, and 2007	Descriptive analysis and one-way analysis of variance (Quantitative)
The *attained curriculum*	6. To what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement? a. At the classroom level, to what extent did the observed changes in OTL levels over time affect changes in aggregated 8th grade students' mathematics achievement between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori*)? b. Using a multilevel modeling approach, to what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement within cohorts? c. Did OTL moderate the relationship between students' socioeconomic backgrounds and mathematics achievement within cohorts, as measured by the TIMSS student assessments?	TIMSS Teacher Questionnaire and Student Questionnaire data from TIMSS 1999, 2003, and 2007	Analysis of covariance and multilevel modeling (Quantitative)

Theoretical Significance

As alluded to earlier in the chapter, OTL is a multidimensional measure, which would be best addressed using a multidimensional OTL conceptual framework like the one proposed by Stevens (1996). OTL measures in Stevens' multidimensional framework have yet to be formally operationalized and validated using the TIMSS data. Given that no empirical studies to date have evaluated all four OTL elements highlighted in Stevens' framework (i.e., content coverage, content exposure, content emphasis, and quality of instruction) simultaneously, this study addressed this gap in the research.

Different OTL levels are likely to result from student membership in different classrooms taught by different teachers located in different schools (Schmidt et al., 2013). Further, students tend to perform better on achievement tests like the TIMSS if they receive more opportunities to learn the tested concepts (Schmidt et al., 2013). Therefore, a good understanding of OTL effects on achievement requires researchers to examine the OTL variables that attend to students' nested-ness within classrooms/schools. To date, research around OTL using multilevel models is limited (Schmidt et al., 2001; Wang, 1998). This is another gap that the present study addressed.

Furthermore, the majority of international comparative research studies has focused on cross-country investigations of OTL and achievement gains (Husén, 1967; Kifer & Burstein, 1992; Schmidt & Burstein, 1993; Schmidt & Maier, 2009; Schmidt, McKnight, Houang, & Wang, 2001; Schmidt et al., 2013). The few within-country OTL studies that exist were conducted by the United States Agency for International Development (USAID) in low-income developing countries such as Ethiopia, Guatemala, Honduras, Mozambique, and Nepal (Adelman, Moore, & Manji, 2011; Cetola, DeStefano, Schuh Moore, & Adelman, 2010; DeStefano, Adelman, & Schuh Moore, 2010; DeStefano & Elaheebocus, 2008; Schuh Moore, DeStefano, & Adelman, 2012). There is a paucity of research on OTL effects related to educational reform efforts in a developed country, such as Japan.

Therefore, this study intended to inform and extend the existing OTL literature as well as offer a fair assessment of the *yutori* reforms over time. By examining the *yutori* reform initiatives at three strategic time points with appropriate qualitative and quantitative data sources, the study shed light on the *intended, implemented,* and *attained curriculum* under *yutori*. The current study further attempted to validate Stevens' conceptualization of OTL as a multidimensional construct using teacher survey data. It used the validated OTL measures to examine the extent to which OTL levels varied in Japanese classrooms as *yutori* reforms were implemented over time. Finally, the study evaluated the relationship of the OTL construct measure taken at the classroom level on achievement measures at the student level. The theoretical significance of the research lies in this in-depth examination of *yutori* reforms, while considering the multidimensionality of the OTL construct and its multilevel influences on students' mathematics achievement.

Policy Significance

Due to the lack of reliable longitudinal data for assessing students' scholastic achievement trends through domestic programs, the entire discourse around *yutori* reforms in Japan was based on results from the two major ILSA programs, namely the PISA and TIMSS. The public and the media have been quick to generalize the crisis of

Japanese education by presenting ILSA results merely in league tables comparing rankings of countries (Takayama, 2008). Unfortunately, the convenience of a single ranking or score to represent the performance of a country's educational system is appealing to policymakers, the media, and the public (Engel & Feuer, 2014). "But, researchers have consistently warned against the inherent dangers in using a single average achievement score as the leading indicator of educational quality" (p. 327). Because of the league table presentation format, ILSA reports tend to be interpreted as a "horse race" between countries (Pizmony-Levy, 2014). According to Braun (2014), using a country's ranking on the ILSAs to set national education goals could do a "disservice to the nation's distinctive culture and educational needs" (pp. 332-333). To minimize negative consequences, unintended or not, researchers have emphasized the importance of making inferences that are supported by ILSA test scores (Feuer, 2013). The *yutori* achievement crisis rhetoric seems to be lacking in terms of the validity of inferences regarding educational quality.

Using the TIMSS test scores, the present research sought to shift the focus to within-country changes in cohort performance over time in the context of educational reforms, using Japan as case. It also considered the influences of OTL in achievement to offer a more valid way of informing the national education policy debate using ILSA data. Because the TIMSS samples are nationally representative groups of schools and students, the findings from this study can be generalized to 8th graders in Japan. No existing research has used a nationally representative sample to study the changes in students' learning opportunities under the *yutori* reforms and how that subsequently affected students' achievement.

This study also sought to add to the literature on curriculum implementation (Fullan, 2007; Sarason, 1971; Spillane, 2004; Tyack & Cuban, 1995) through a detailed examination of *yutori* reform implementation. Unlike similar U.S. reform efforts such as the National Council of Teachers of Mathematics (NCTM) Standards 2000, *yutori* reforms actually outlined concrete curricular changes—a cut in instructional time and introduction of IS classes—to the elementary school and the junior high school curricula. This dissertation endeavored to inform other future reform efforts by shedding light on the implementation processes of *yutori* reforms.

Definition of Terms

Definitions of terms salient to this study are provided below to facilitate the interpretation of the literature review, procedures, and findings.

- *Opportunity to Learn* (OTL). The opportunity for a student to learn concepts and skills emphasized through organized curricula at school. The OTL definition used in this study was based on Stevens (1993) and has four dimensions:
    - *Content Coverage*—The core curriculum topics covered specific to a particular grade level or subject area.
    - *Content Exposure*—The amount of time teachers allocated to covering the content and depth of the teaching provided.
    - *Content Emphasis*—The emphasis given to certain topics that are part of the core curriculum.

  - *Quality of Instructional Delivery*—The coherence and effectiveness of classroom teaching practices.

- *TIMSS Curriculum Model*. The tripartite model that contains three curriculum aspects (Foy, Arora, & Stanco, 2013; Foy & Olson, 2009; Gonzalez & Miles, 2001; Martin, 2005):
  - *Intended Curriculum*—The curriculum set at the national or system level (school district, regions, etc.) that represents the nation's or system's objectives and traditions. The *intended curriculum* is what the governing body of education experts expects to be taught in classrooms and answers the question, "What are students expected to learn?"
  - *Implemented Curriculum*—The *implemented curriculum* is comprised of what is actually taught to students. The classroom teacher is the central agent delivering the mathematics curriculum, and the choices that the teacher makes in terms of instructional materials, emphasis on particular topics, and other teaching practices all have fundamental implications for the implemented curriculum. It addresses the question, "How is instruction organized?"
  - *Attained Curriculum*—The *attained curriculum* refers to the new mathematics knowledge that the student has acquired as a result of being taught the curriculum in school. It answers the question, "What have students learned?

Chapter II

REVIEW OF THE LITERATURE

This chapter provides a review of the literature to address the historical context of the *yutori* reforms, the Japanese education system, and Opportunity to Learn (OTL). In this particular order, it discusses the following: (a) a historical background of the Japanese education system, (b) a description of the components of the *yutori* reforms, (c) a review of research on OTL, (d) the OTL theoretical framework adopted and the gaps/issues that this study addresses, and (e) a review of research on the relationship between OTL and student achievement in mathematics.

Historical Background of the Japanese Education System

Japan has been commonly perceived as an egalitarian society but with a meritocratic philosophy and educational practices (Kariya, 2009; Wilkinson & Pickett, 2011). The egalitarianism and equality values are supported by the conceptualization of the education system as a mechanism of meritocratic achievement (Kariya, 2009). The Japanese school system is based on a meritocracy: a social system that offers educational opportunity as a function of merit; merit is determined by academic performance in school as recorded on exams (OECD, 2012; Rohlen, 1983, 1988).

Education is compulsory in Japan at the primary and lower secondary school levels, that is, Grades 1 through 9. After 9th grade, Japanese students take an examination to get into high school and again to gain admission into college. The high school and university entrance exams represent gateways to status in Japanese society (Tsuneyoshi, 2004). Exam success reflects not only on the individual, but also on the students' families and teachers. This support group shares the responsibility for failure and creates the pressure to succeed because the emphasis on *where* a person studied, rather than *what* they studied, is strong in Japan (OECD, 2012).

Since as far back as 1964, with the First International Mathematics Study (FIMS), Japanese students have performed well compared to other nations. Japan has consistently been placed at or near the top of the rankings based on student results on various international large-scale assessment (ILSA) programs. Japan also has the highest secondary school graduation rate (95%) among the Group of Eight (G8) nations (OECD, 2012). Japan has consistently achieved high performance ratings by anticipating the changing demand for skills and competencies rather than simply reacting to them (OECD, 2012).

The path leading to *yutori* reforms began in the 1970s, when Japan was rebuilding after the Second World War. In 1971, OECD published an advisory report suggesting the Japanese government was paying more attention to "the development of students' personalities through a more flexible and less pressured scheme of education" (as cited in Bjork, 2009, p. 29). In the same year, the Central Council for Education (CCE) also announced a set of reform guidelines that coincided with the OECD's recommendations. These were: (a) a more flexible curriculum, (b) increased emphasis on personal expression and internationalization, and (c) experimentation with instructional methodology (CCE, 1972). Many issues highlighted by the CCE continued to undergird the Japanese Education Ministry's (Ministry of Education, Sports, Culture and Technology or MEXT) current goals.

In the early 1980s, the MEXT expressed growing concerns over intensified academic pressure which has resulted in some students being left behind. Many children had no future aspirations and little desire to progress in school, as evidenced by the increase in the number of young people who were not engaged in any forms of employment or education. A 2012 OECD report stated that there is "too much competition to get into good schools and parents lack the confidence in formal public schooling…" (p. 184). Concurrently, the pressure to succeed also contributed to a variety of social problems within schools, including bullying, school violence, and suicide (Bjork, 2015; Cave, 2003; Kariya, 2009).

In the 1990s, Japan underwent a recession that brought about a prolonged economic slump for the next decade. Class differences were beginning to crack the longstanding societal base of egalitarianism. The Japanese education system was criticized for producing graduates with strong basic academic skills but little creativity and independence. There was too much emphasis on inculcating knowledge for the sake of examinations, yet not enough attention paid to encouraging independent, self-motivated enquiry (Cave, 2003). The Japanese business community asserted that Japan's global economic position was closely tied to the cultivation of originality, thinking ability, and diverse learning experiences; therefore, "creativity should be ranked as the most desirable qualification [among students today]" (Japan Committee for Economic, 1984, p. 35). This discourse was, in large part, a reaction to the uniformity of the curriculum and the rigors of learning with a strong focus on standardized examinations at the expense of developing creativity and individual inquiry among its students (Kariya & Rappleye, 2010). The CCE issued reports that rejected the old forms of Japanese education, instead calling for an emphasis on greater freedom for the child in the schooling process, including more exploratory and experience-driven teaching methods (CCE, 1997). This push for creativity and self-expression only grew stronger when the bubble economy years ended.

Emergence of the *Yutori* Reforms

In response to these criticisms, MEXT reduced the number of school days in 1992—removing one Saturday per month from the school calendar. Then 4 years later, MEXT proposed a new education initiative, "*zest for living,*" as the main objective of education with the fundamental premise of nurturing students' self-learning competencies for critical discovery and problem-solving skills in a rapidly changing society (CCE, 1997; OECD, 2009). In December 1998, MEXT announced a new education policy, as part of the *zest for living* initiative, for kindergartens, primary schools, and junior high schools; this was eventually implemented in 2002. This reform was called "*yutori-kyoiku*" or "*relaxed education,*" which denoted that students need a liberal, flexible, and comfortable school life to develop their individuality (MEXT, 2000).

Yutori curricular reforms constituted a 30% cut in the curriculum of core academic subjects, the implementation of a 5-day school week instead of a 6-day school week, and the introduction of an "*integrated studies*" (IS) dedicated to student-centered, experiential learning focused on key competencies of *zest for living* in elective subjects (MEXT, 2002). The 5-day school week was, in fact, phased in gradually since 1992. Up until 2002, two Saturdays a month had been holidays. By eliminating all Saturday classes through the *yutori* reforms, this meant a cut of over 10% of class hours, which gave Japanese children more free time than before.

The IS classes aimed to cultivate students' ability to discover questions and the ability to learn, think, and decide independently (MEXT, 2002). They were first introduced in the FY1998 revision of the Course of Study, a guideline for national curriculum (MEXT, 2000). Schools could select topics for IS that match the interests of their students and the unique characteristics of their communities. In other words, through the reforms, the Japanese government allowed greater teacher autonomy in instructional approaches, while implicitly encouraging teachers to use student-centered methods to stimulate critical thinking and creativity. However, no designated instructors were hired to teach IS classes. Instead, teachers from all disciplines were expected to take part in the planning and implementation of IS classes (Bjork, 2015).

MEXT deliberately kept its directives and reform guidelines for *yutori* reforms at a minimum. Unlike the detailed, prescribed curriculum provided to teachers in the past, the new *yutori* curriculum offered very little guidance. As Peter Cave (2003) reported, "In the new junior high curriculum, nine pages are given to Japanese and mathematics, fifteen to science, and nineteen to social studies. In contrast, only one page is devoted to Integrated Learning" (p. 90).

Student Performance Under *Yutori* Reforms

The new reform efforts generated numerous and heavy criticism even before they were fully implemented (Tsuneyoshi, 2004). Critics from diverse sectors began to link the *yutori* reforms to the lowering of academic standards as Japanese students were required to study less than before, or at least less than their peers in other industrialized countries (Kariya & Shimizu, 2004; Stevenson & Stigler, 1992). Most importantly, the public raised concerns that the new cut in curriculum content in core subjects would lead to a drop in measured knowledge levels (Cave, 2003). Many believed that the Ministry of Education was improving creativity and fostering independent learning at the expense of the deterioration of knowledge. The public debates framed *yutori* reforms as having an impact on the nation's education crisis: "If 30% of the curriculum content was reduced and schools hours were reduced as well,…it seems common sense to assume that achievement will suffer" (Tsuneyoshi, 2004, p. 388).

Soon after the implementation of *yutori* reforms, the release of the PISA 2003 and TIMSS 2003 results sent shockwaves throughout Japan. Due to a lack of reliable longitudinal data for assessing the students' scholastic achievement trends through domestic programs, the media, the public, and the policymakers used the ILSA results extensively to criticize *yutori* reforms (Takayama, 2007, 2008; Tsuneyoshi, 2004). On international league tables, the average performance of Japanese 15-year-olds dropped from 1st to 6th rank in mathematics and from 8th to 14th in reading on the PISA 2003 tests. Compared to their performance on the PISA 2000, the Japanese students' mean performance on PISA 2003 assessments dropped from 557 to 543 in mathematics, from 522 to 498 in reading, and from 550 to 548 in science. In particular, the Japanese reading literacy scores on the PISA 2003 test were statistically significantly lower than their scores on the PISA 2000 test (OECD, 2004; Takayama, 2008).

Between TIMSS 1999 and TIMSS 2003, there was also a decline in the mean mathematics (from 597 to 565 among Japanese 4th graders; from 579 to 570 among Japanese 8th graders) and science test scores (from 574 to 543 among Japanese 4th graders). Further, Japanese students "reported the lowest interest in and enjoyment in mathematics" among all participating countries of PISA 2003 (OECD, 2004, p. 119). Similarly, only 9% of the 8th graders tested on the TIMSS 2003 tests strongly agreed with the statement "*I enjoy learning mathematics*," compared to the corresponding international average of 29% (Mullis et al., 2004, pp. 159-160).

However, the "crisis" reporting of PISA 2003 and TIMSS 2003 results suggesting a decline in the mean mathematics test scores and ranking may have been unfounded. The top-performing nations, Hong Kong and Japan, were not statistically different from each other on mean mathematical literacy scores (OECD, 2004). Moreover, there were no statistically significant changes in Japanese 15-year-olds' mathematical literacy levels between administrations of the PISA 2000 to PISA 2003 tests (see OECD, 2004).

The findings of low learning motivation levels in students based on self-report surveys, observed on the PISA 2003 and TIMSS 2003 testing, were nothing new for the nation. Japanese students had historically exhibited higher academic performance and lower motivation levels for learning mathematics on surveys, as compared to other nations, since the FIMS (Takayama, 2008).

Barriers to Implementing *Yutori* Reforms

One notable feature of the *yutori* reforms was the significant amount of autonomy granted to local actors (Bjork, 2009). "MEXT delivered reform guidelines to the schools but entrusted them to devise concrete strategies for realizing those plans" (p. 31). In other words, the government did not provide schools and teachers with adequate resources and support to roll out the *yutori* reform effort (Bjork & Tsuneyoshi, 2005; Kariya & Rappleye, 2010). Teachers were not given the time and necessary training in new pedagogical approaches, classroom management, and lesson design to realize this ambitious policy vision (Bjork, 2009; Cave, 2003; Wada & Burnett, 2011).

Another deficiency of the *yutori* reforms was MEXT's vague articulation of expected policy outcomes (Bjork, 2015). There was an implicit expectation that teachers and schools had the skills and commitments to improve student learning as a part of *yutori* reforms. Specifically, MEXT encouraged schools to "evaluate the learning status of each and every child even more carefully than before" (MEXT, 2003, p. 24). Schools and teachers were entrusted with assessment procedures. Absent from *yutori* reform policy documents was any mention of how the outcomes of the reforms would be measured (Bjork, 2015).

Lastly, the reforms also appeared to exacerbate the inequality in the academic achievement levels of students from low- and high-income families (Kariya, 2010; Kariya & Shimizu, 2004; Mimiduka, 2007). Despite efforts made to relieve academic pressures experienced by students, corresponding changes were not made to the high school and university entrance examination system (Wada & Burnett, 2011). In a 2008 survey by MEXT, more than half of the parents indicated that "school classes alone were not sufficient," and another 14% believed that "school classes alone could not adequately prepare children for school entrance exams" (MEXT, 2008). Such dissatisfaction with schools motivated parents to rely on *juku* (private, afterschool tutoring institutions). This also widened the gap between students whose families could afford to send them to *juku* and those who could not. In other words, the reforms further exacerbated social stratification tied to inequality in educational achievement levels (Kariya, 2010; Kariya & Shimizu, 2004; Mimiduka, 2007).

Literature Review of Opportunity to Learn (OTL)

What Is OTL?

Opportunity to Learn (OTL) generally refers to the opportunity for a student to learn important concepts and skills. This concept is fundamental to schools, which are organized to provide students with the time and experiences geared towards learning specific types of subject matter. As early as the turn of the last century, OTL was mentioned in the writings of Edward L. Thorndike and William James (see Berliner, 1990). Starting in the 1960s, separate OTL research strands started to emerge around three different aspects of instruction: time on instruction (Carroll, 1963), content of instruction (Husén, 1967), and quality of instruction (Brophy & Good, 1986). Finally, a multidimensional OTL framework was proposed in the early 1990s (Stevens, 1993). The three research strands and the multidimensional OTL framework are presented below.

OTL as Time on Instruction

Carroll's model. The notion of OTL was first introduced by John B. Carroll in his seminal model of school learning, which extended the generic "educational opportunity" from a "yes or no" dichotomy to a continuum expressed as the time allowed for learning (Carroll, 1963). Carroll argued that anyone could succeed in learning a given task as long as he or she spends the needed time. The question was no longer "*What* can this person learn?" but "*How long* will it take this person to learn?" Carroll hypothesized that the degree of student learning or the amount of time needed is a function of five factors (Carroll, 1989): (a) *Aptitude*: the amount of time a learner needs to learn a given task under optimal instructional conditions; (b) *Ability to understand instruction*: the ability of a learner to understand what the learning task is and how to go about learning it; (c) *Perseverance*: the amount of time the learner is willing to spend on learning the task; (d) *Opportunity to learn*: the amount of time allocated to learning a concept; and (e) *Quality of instruction*: the degree to which instruction is presented so as not to require additional time for mastery beyond that required by the aptitude of the learner.

The first three factors are internal characteristics of the student while the last two factors are external to the student, which can be potentially shaped by teachers, schools, and other aspects of the education system and outside. The model can be expressed in the metric of time as the following:

$$Degree\ of\ Learning = f\left(\frac{Time\ actually\ spent\ learning}{Time\ needed\ to\ learn}\right)$$

According to the model, the degree of learning is a function of the ratio between time spent learning and time needed to learn. The numerator "*Time actually spent learning*" is composed of "*Opportunity to learn*" and "*Perseverance*"; the denominator "*Time needed to learn*" is composed of "*Aptitude,*" "*Quality of instruction,*" and "*Ability to understand instruction.*" The full Carroll model is presented below:

$$Degree\ of\ Learning = f\left[\frac{\begin{pmatrix}Opportunity\ to\ learn\\ or\\ Time\ allocated\\ for\ learning\end{pmatrix}\times\begin{pmatrix}Perseverance\\ or\\ Percentage\ of\ time\\ actually\ spent\\ engaged\ in\ learning\end{pmatrix}}{\begin{pmatrix}Aptitude\\ or\\ Time\ actually\\ needed\ to\ learn\end{pmatrix}\times\begin{pmatrix}Quality\ of\\ instruction\end{pmatrix}\times\begin{pmatrix}Ability\ to\\ understand\\ instruction\end{pmatrix}}\right]$$

With this model, Carroll turned OTL into an instructional time concept: degree of learning depends on amount of time allocated for learning. It is intuitive that unless a student was provided with the opportunity to learn some things, he or she might not learn them.

Bloom's mastery learning model. Benjamin Bloom (1968) later elaborated Carroll's model into a working model for mastery of learning. Bloom argued that by adjusting the instructional variables in Carroll's model—namely, OTL and quality of instruction—any student could achieve some mastery performance level if attention were paid to increasing the time spent or decreasing the time needed to learn or both (Bloom, 1974). Further, Bloom (1974) contended that "if teachers and curriculum makers can define an appropriate criterion of achievement, it then becomes the responsibility of the teachers and schools to provide the time necessary for students to attain the criterion" (p. 683). In other words, given sufficient time and appropriate instruction, virtually "*all students could learn well.*" The mastery learning model offers a broader, theoretical basis for understanding the instructional process and explaining school learning effects, which in turn becomes a practical way for educators to enhance achievement.

Wiley and Harnischfeger's model of instructional exposure and achievement. Using Carroll and Bloom's models as theoretical bases, Wiley and Harnischfeger (1974) made a strong case for understanding schooling from the students' perspective. They argued that the amount of schooling students partake of must be mediated through the students' pursuit. Achievement comes from the active behavior of students—their involvement in their own learning. Further, all the educational variables (i.e., teachers' skill, curriculum material, allocated time policies, etc.) could affect achievement only through the amount of time students spent actively engaged in learning. Therefore, according to Harnischfeger and Wiley (1985), the primary way to understand how schools accomplish their instructional goals is to study *what* students attend to and the *duration* of that attention.

Academic learning time. Following the lead of Harnischfeger and Wiley, the project staff of the Beginning Teacher Evaluation Study (BTES: Berliner, 1990) developed their own model of academic learning time to help them understand classroom instruction (Berliner, 1978, 1990; Denham & Lieberman, 1980). Their model is called "Academic Learning Time" (ALT), which defined OTL as the amount of time a student spends engaged with materials and activities in which a high level of success is attained, and in which the materials and activities are related to outcomes that are valued (Fisher et al., 1981). Four variables that make up ALT are: (a) allocated time, (b) engaged time, (c) success rate, and (d) the degree of alignment of the curriculum with the outcome measure.

The ALT model differs from Carroll's model in two ways. First, the ALT model includes the curriculum content areas and the outcome measures to assess effects of that curriculum content. It recognizes that even time-on-task is not sufficient to measure learning outcomes; what is really necessary is a "time-on-the-right-task" measure

(Berliner, 1990). For a student to attain a particular achievement standard, the student needs to be provided with not only enough time to do so, but also the time needed to be engaged with a curriculum that is logically related to the standard (Berliner, 1978). That is, the opportunity that counts is one in which the student is paying attention—and paying attention to material aligned with the intended outcome measures.

A second distinction is the inclusion of success rate in the ALT model to provide a quantifiable time metric for the two non-time variables in the Carroll model—*quality of instruction* and *ability to understand instruction*. The ALT model used the following logic to operationalize these concepts: If success rate for a student is high, then either the quality of instruction or the ability to understand instruction, or both, must be high. Conversely, if a student's success rate is low, then either the quality of instruction or the student's ability to understand instruction, or both, must be low.

In sum, the teacher's role in determining students' opportunities or time allocated in learning a specific topic is crucial for both Wiley and Harnischfeger's notions of OTL and the ALT model. Teachers could allocate time to various topics as they saw fit. However, this allocated time is reduced by the amount of time the teacher engaged in classroom management tasks (e.g., discipline, maintaining order in the classroom, etc.). Therefore, Wiley and Harnischfeger reasoned that a student's OTL is heavily dependent on the length of the school day and school year, but also on individual teachers within the school.

OTL as Content of Instruction

The International Association for the Evaluation of Educational Achievement (IEA) model. A second OTL research strand focusing on content overlap between the enacted and assessed curriculum emerged with the IEA studies. In the late

1950s, the interest in exploring the variables related to school effectiveness and student learning that could be compared across school systems sparked the creation of the Council of the International Association for the Evaluation of Educational Achievement. The council consisted of scholars, educators, psychologists, and psychometricians, including Benjamin Bloom. In 1964, these IEA scholars conducted the First International Mathematics Study (FIMS) in 12 countries to assess 13-year-old and final-year secondary students' mathematics achievement (Husén, 1967). The primary objective of FIMS was to investigate various school systems by examining the influences of various school inputs on students' achievement scores in the participating countries (Husén, 1967). Bloom, Husén, and other FIMS researchers contended that one of the factors which may influence scores on the achievement examination was "whether or not students have had the opportunity to study a particular topic or learn how to solve a particular type of problem presented by the test" (Husén, 1967, pp. 162-163).

Up until that time, the OTL construct had been conceived as a time-based variable operating at the individual student level. This presented a challenge of measuring the OTL construct at the classroom or teacher level through a large-scale survey. This also presented a problem when comparing countries as they have different national curricula or educational systems; variations in what content is covered are bound to occur. Building on Carroll and Bloom's work, the FIMS researchers developed another way of conceptualizing OTL—as a content-based variable focusing on the teaching-learning process that occurs in schools. FIMS investigators included OTL as a measure of teachers' perception of students' opportunity to become familiarized with the material covered by the test item (Schmidt & Maier, 2009).

Specifically, for each test item, teachers were asked to indicate the proportion of their students who have had an opportunity to learn that particular type of problem. This conception of OTL focuses on the content covered rather than on the time allocated in Carroll's model. Including OTL in FIMS led to two important findings (Husén, 1967): (a) there was a significant positive association between the teachers' assessments of opportunity to learn and the mathematics scores, and (b) a considerable amount of between-country difference in mathematics scores could be attributed to the differences between students' opportunities to learn the material that was tested.

In the same spirit, OTL had a more pivotal role in the Second International Mathematics Study (SIMS), which was conducted by IEA in early 1980s. With its main objective centering on the curriculum, SIMS introduced a three-tier model of the curriculum: *intended*, *implemented*, and *attained* (Travers & Westbury, 1989). This model viewed the mathematics curriculum in a school system as having three aspects, each associated with a different level of the system.

This model has continued to underpin IEA studies in mathematics and science to this day. *The intended curriculum* is the curriculum set at the national or system level (school district, regions, etc.) that represents the nation's or system's objectives and traditions (Mullis et al., 2009; Travers & Westbury, 1989). These goals are often articulated through official documents such as national curriculum guides, course syllabi, and prescribed textbooks. The intended curriculum is what the governing body of education experts expects to be taught in classrooms and answers the question, "What are students expected to learn?"

At the school or classroom level, *the implemented curriculum* includes what is actually taught to students (Mullis et al., 2009; Travers & Westbury, 1989). The classroom teacher is the central agent in delivering the mathematics curriculum, and the choices that the teacher makes in terms of instructional materials, emphasis on particular topics, and other teaching practices all have fundamental implication for *the implemented curriculum*. It addresses the question, "How is instruction organized?"

At the individual student level, *the attained curriculum* refers to the new mathematics knowledge that the student has acquired as a result of being taught the curriculum in school (Mullis et al., 2009; Travers & Westbury, 1989). It answers the question, "What have students learned?" and may be considered the final outcome of the educational process as represented by the students' test scores. The match between the *intended*, *implemented*, and *attained* curricula was an important focus of SIMS.

OTL was operationalized as the implemented curriculum or implemented coverage in SIMS (Travers, Garden, & Rosier, 1989; Travers & Westbury, 1989). The original FIMS OTL question asked teachers to estimate the percentage of their students who had an opportunity to learn a particular mathematics item on a 3-point response scales ranging from *All or most (at least 75%)* to *Few or none (under 25%)*. For SIMS, this mathematics OTL question was replace by a pair of questions:

1. What percentage of the students from the target class do you estimate will get the item correct without guessing?
2. During this school year, did you teach or review the mathematics needed to answer the item correctly? (Floden, 2002, p. 11)

Asking this pair of questions allowed the teacher to indicate whether the mathematics related to the test item had been taught and the approximate percentage of students in the class who would likely get the question correct, accounting for individual variability and measurement error (Floden, 2002).

The refined conception of OTL helped SIMS researchers gain a better understanding of the between-country variation in OTL (Suter, 2017). For instance, Japan and France were found to demonstrate relatively homogeneous OTL ratings. The United States displayed considerably more variation (Schmidt, Wolfe, & Kifer, 1992).

A within-country perspective on OTL provides a basis for considering current practice and possible policy alternatives. For instance, Schmidt and colleagues (2001) found that ability grouping or tracking in eighth-grade mathematics classes contributed to variation in OTL ratings on the teacher surveys in the United States. The students in tracked classes had higher ratings on opportunity to learn, whereas those in non-tracked classes had lower opportunity to learn ratings; such differences led to different student achievement levels on SIMS mathematics outcome measures.

Though substantial differences in OTL between and within (some) countries were found on SIMS, the study did not detect large effects of OTL on students' mathematics achievement. This puzzling finding implied that the employed measures of OTL on SIMS may have been inadequate, given what was known about various levels of curricula (Schmidt & Maier, 2009). Another major criticism of the OTL measures on SIMS is that "it is too bound to the form of specific items and [are] more representative of teachers' judgments of items rather than content categories of which the item is an example" (Schmidt & McKnight, 1995, p. 345).

In the 1990s, the National Science Foundation (NSF) and the National Center for Education Statistics (NCES) jointly funded the Survey of Mathematics and Science Opportunities (SMSO) to develop and validate instruments for the IEA Third International Mathematics and Science Study (TIMSS), with a special focus on advancing and operationalizing the OTL construct (Schmidt & McKnight, 1995). SMSO proposed several changes to OTL measures in the TIMSS. Figure 2 presents the new TIMSS framework on educational opportunity.

At the national or system level, the TIMSS gleaned information on *the intended curriculum* from official curricular documents and used that information to ask officials from the Ministry of Education or influential educators from each country to detail the actual content that was intended to be covered at each grade level (Schmidt & Maier, 2009).

The second change to the OTL measure involved *the implemented curriculum.* In the FIMS and SIMS, teachers were asked for an item-by-item rating of whether content sufficient to answering the item had been taught to students. This task was difficult for teachers because "teachers had to view [too] many items and it required teachers to abstract content related to performance from each particular item" (Schmidt & Maier, 2009, p. 545). To resolve this problem, the TIMSS provided a list of topics, based on the curricular frameworks developed from analyses of *the intended curriculum*, for each grade level. The measurement of OTL in the TIMSS was done by "naming a topic, giving more than one item to illustrate the topic, then asking the teacher about opportunity to learn (in terms of) similar exercises that address this topic" (Floden, 2002, p. 239). Therefore, teachers were encouraged to think about OTL with reference to a given topic rather than a specific item.

Lastly, the TIMSS refined *the attained curriculum* by tying assessment more closely to the curricular framework, which yielded "curriculum sensitive OTL measurement." A series of subtest scores, specific to the content strata of the curriculum, was developed to yield scores that were aligned with school instruction. For example, TIMSS mathematics assessment included six major content areas at the 8th grade. Consequently, test results at each subtest-level were available for each country on the topics of: (a) fractions and number sense; (b) geometry; (c) algebra; (d) data representation, analysis, and probability; (e) measurement; and (f) proportionality (Schmidt, McKnight, Cogan, & Jakwerth, 1999). This resulted in performance that varied by country.

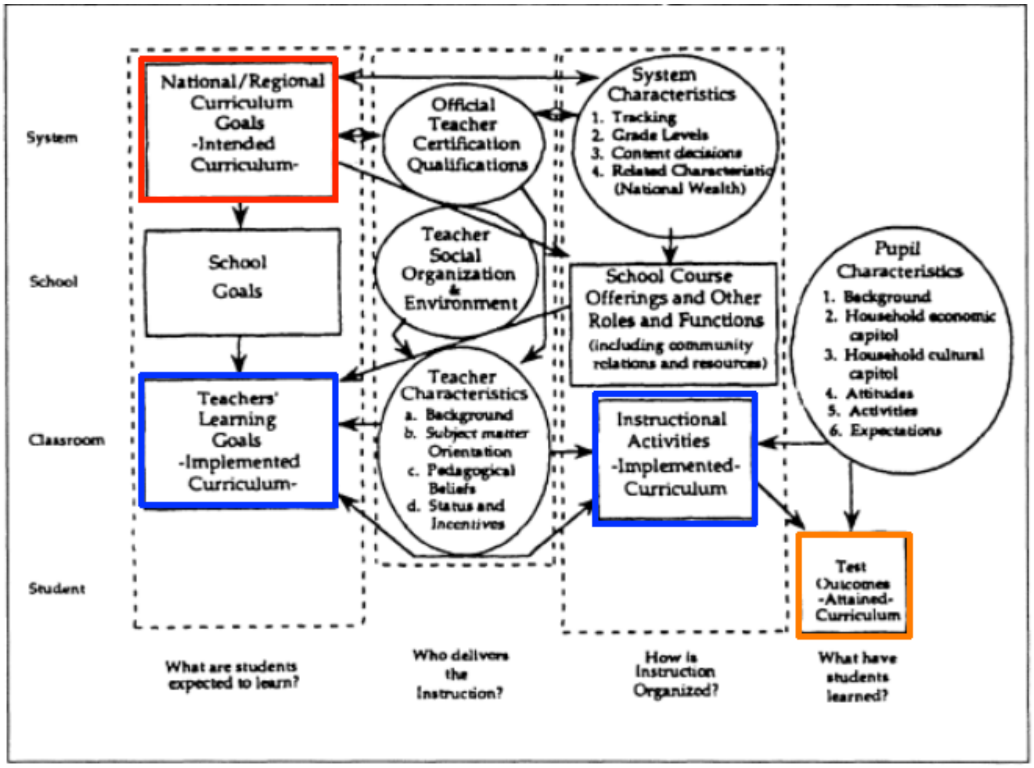

System
School
Classroom
Student
National/Regional Curriculum Goals -Intended Curriculum-
School Goals
Teachers' Learning Goals -Implemented Curriculum-
Official Teacher Certification Qualifications
Teacher Social Organization & Environment
Teacher Characteristics
a. Background
b. Subject matter Orientation
c. Pedagogical Beliefs
d. Status and Incentives
System Characteristics
1. Tracking
2. Grade Levels
3. Content decisions
4. Related Characteristic (National Wealth)
School Course Offerings and Other Roles and Functions (including community relations and resources)
Instructional Activities -Implemented- Curriculum
Pupil Characteristics
1. Background
2. Household economic capitol
3. Household cultural capitol
4. Attitudes
5. Activities
6. Expectations
Test Outcomes -Attained- Curriculum
What are students expected to learn?
Who delivers the Instruction?
How is Instruction Organized?
What have students learned?

Figure 2. TIMSS Conceptual framework: Educational opportunity
Adapted from "Surveying educational opportunity in mathematics and science: An international perspective" by W. Schmidt & C. McKnight, 1995, *Educational Evaluation and Policy Analysis, 17*(3), p. 349.
Copyright 1995 by the American Educational Research Association.

The Programme for International Student Assessment (PISA) approach. In 2012, another notable international student assessment program, the PISA, added questions addressing the OTL construct in the mathematics portion of its assessment. The PISA is a worldwide study conducted in member and non-member nations and sponsored by the Organisation for Economic Co-operation and Development (OECD). It surveys 15-year-old school pupils' scholastic performance on mathematics, science, and reading tests.

The operationalization of OTL in the PISA assessments is mainly based on student judgments (OECD, 2012). The rationale for students reporting their own OTL is a function of the PISA's age-based rather than grade-based methodology.

Different from the TIMSS framework which focuses on what students know after studying particular grade-level curricula over a period of time, the PISA framework aims to "assess students' ability to use what they have learned through their accumulated schooling experience to address real-life challenges" in a non-graded manner (Cogan & Schmidt, 2015, p. 210). The PISA differs from the TIMSS in both the definition of the student population and the sampling methodology. Rather than sampling intact 8th grade classrooms, the PISA randomly samples 15-year-old students from all classes in schools. Additionally, the mathematical assessment portion of the PISA is not framed according to curricular elements but refers to fundamental abilities in four topical areas: *quantity*; *uncertainty and data*; *change and relationships*; and *space and shape*.

The PISA 2012 operationalized OTL as students' familiarity with and exposure to a small set of key mathematics topics typically found in Grades 8 through 12 as well as real-world applications and word problems. Students were presented with a series of mathematics tasks identified in the PISA mathematics framework. Following each of the questions, students were asked to judge "whether and how often they have seen similar task in their mathematics classes and in previous assessments" (OECD, 2012, p. 187). Three OTL indices indicating the frequency with which students encountered specific topics/situations were then derived from these student questionnaire items: (a) OTL related to formal mathematics, (b) OTL related to applied mathematics, and (c) OTL related to word problems. Each of these indices could have values ranging from 0 (never) to 3 (frequently). These student-level OTL measures could be aggregated and analyzed at the school- and country-level countries, according to OECD (Cogan & Schmidt, 2015; OECD, 2012; Schmidt & Burroughs, 2013).

Although the three indices provide a chance to examine the relationship between schooling and mathematics literacy as defined by the PISA, these OTL items are not as extensive as the ones asked in TIMSS. A major weakness of the PISA OTL data is that they lack classroom-specific OTL information from teachers and courses (Schmidt et al., 2013). As such, variations in student OTL could easily be confounded. For instance, the specific opportunities students have experienced individually could be due to their different course-takings. Even if students took the same course, their experience would differ, depending on their teachers (Schmidt et al., 2013).

The Content Determinants Project. Another line of content of instruction research focused on students' opportunity to learn important content objectives. Andrew Porter and colleagues started the Content Determinants Project in the Institute for Research on Teaching (IRT) to study how teachers make curricular decisions that have important consequences for students' OTL (Floden, Porter, Schmidt, Freeman, &

Schwille, 1981). These IRT researchers contended that teachers are presented with multiple statements—those coming from standardized tests, textbooks, and administrators—of what content should be included in a particular grade level for any academic subject. To study teacher decision making, both the content actually covered by teachers and the content implicit in the many curricular statements aimed at teachers must be examined jointly. Porter and colleagues later shifted their work away from content coverage and toward alignment. They not only examined classroom-level content coverage or implied content coverage expected via national standards, but also examined the cognitive demand associated with content coverage (Gamoran, Smithson, & White, 1997; Porter, 2002; Porter & Smithson, 2001). Following such logic, Porter and colleagues defined OTL as a combination of topics and cognitive demand. They argued that student achievement on a test is dependent on the alignment between the content covered by OTL and the content covered by the assessment. This alignment measure, now called the Survey of Enacted Curriculum (SEC), assesses alignment between intended, enacted, and assessed curricula along a *content topics* and *cognitive demand* matrix.

OTL as Quality of Instruction

The third OTL research strand—quality of instruction—could be traced as far back as Carroll’s (1963) model of school learning. Carroll’s model and later Herbert Walberg’s (1980) model of educational productivity both featured the quality of instruction variable alongside instructional time. In their review of correlational and experimental research done in K-12 classrooms during 1973-1983, Brophy and Good (1986) identified aspects of giving information, questioning students, and providing feedback as important instructional quality factors with consistent empirical support. In a pilot study for a statewide assessment of middle school mathematics, Herman, Klein, and Abedi (2000) operationalized four commonly identified OTL constructs in the literature: curriculum content, instructional strategies, quality instructional resources, and general preparation for the assessment. They performed multitrait-multimethod (MTMM) analysis to examine the construct validity of teacher- and student-reported data on these four measures. Their MTMM analyses found reasonable reliabilities (alpha coefficients ranging from 0.56 to 0.86) and high correlations between teacher- and student-reported data within the two quality of instruction OTL measures (homotrait-heteromethod correlations ranging from 0.52 and 0.53). Herman et al.'s (2000) MTMM analyses results offered one of the first validated OTL instruments. Thereafter, researchers further considered instructional strategies and instructional resources to be crucial aspects of the quality of instruction in OTL operationalization.

Multidimensional OTL Framework

From the 1990s, school accountability reforms in the United States intensified the discussion on the measurement of the OTL conceptual framework (Brewer & Stasz, 1996; Guiton & Oakes, 1995; McDonnell, 1995; Porter, 1991, 1993, Stevens, 1993, 1996; Wang, 1998). Researchers and policymakers proposed to broaden the OTL concept to represent schooling processes and classroom activities so that schools and teachers could have a vision of how to improve (McDonnell, 1995; Stevens, 1996). From her review of a series of international and national research studies, Stevens (1993, 1996) provided the first multidimensional conceptual framework of OTL, identifying four OTL elements that have been prevalent in research: *content coverage, content exposure, content emphasis,* and *quality of instructional delivery.*

Stevens identified a critical gap in the OTL research, namely that most OTL studies “looked at one variable at a time” (p. 4). Stevens’ framework offered a way to examine the aforementioned three OTL research strands from multidimensional perspectives. Most importantly, Stevens (1993, 1996) treated OTL as a teacher effect related to the allocation of adequate instructional time covering a core curriculum via different cognitive demands and instructional practices that could produce student achievement. Kurz, Elliott, Kettler, and Yel (2014) also stressed the same point: “To provide OTL, a teacher must dedicate instructional time to covering the content prescribed by the intended curriculum using pedagogical approaches that address a range of cognitive processes, instructional practices, and grouping formats” (p. 162). The four elements of Stevens’ multidimensional OTL framework are described below.

Content coverage refers to the core curriculum topics covered specific to a particular grade level or subject area. Sample *content coverage* questions include: how many of the items on the test match the curriculum that was taught (Leinhardt & Seewald, 1981; Walker & Schaffarzick, 1974) and do all students have access to the core curriculum (Wiley, 1990; Yoon, Burstein, Gold, Chen, & Kim, 1990). This is the most frequently studied OTL dimension and has been measured in various ways, including teachers’ self-report, direct observations, and analysis of the curriculum materials. *Content coverage* is often measured in three ways: teacher’s self-reports, direct observation of classroom instruction, or analysis of the content of curriculum materials (Winfield, 1993). Past research has found that teachers’ self-reports are reliable indicators of content coverage (Leinhardt & Seewald, 1981; Yoon et al., 1990).

Content exposure refers to the amount of time teachers allocate to covering the content and the depth of the teaching provided. This dimension of OTL can be measured through the time allotted to students to learn (i.e., time on task), the time devoted to a certain subject area, and the amount of time in class periods (Brophy & Good, 1986; Stedman, 1994; Wang, 1998; Wiley, 1990; Winfield, 1987).

Content emphasis refers to the emphasis given to certain topics that are part of the core curriculum. It concerns the issue of which topic within the core curriculum is treated as a major topic, a minor review, or not taught at all (Floden et al., 1981; Goldenberg & Gallimore, 1991; McDonnell, Burstein, Catterall, Ormseth, & Moody, 1990; Shavelson & Stern, 1981; Wang, 1998). This variable also concerns the curriculum offerings differentiated according to student ability levels (i.e., ability grouping and tracking). Students in different tracks were paced differently: students in lower-ability classrooms paced more slowly than students in higher-ability classrooms (Schmidt & Maier, 2009). Tracking differences in either the quantity or quality of education may influence differences in student achievement. Teachers choose what they want to emphasize based on their personal experiences; professional experiences; perception of certain topics as important; and influence of past teachers, courses, textbooks, and other authorities (Floden et al., 1981). A variety of tools have been used to address this component of OTL, including teacher surveys, content analyses of instructional materials, and teacher interviews (Floden et al., 1981; Goldenberg & Gallimore, 1991; McDonnell et al., 1990; Shavelson & Stern, 1981; Wang, 1998).

Quality of instructional delivery refers to how coherently and effectively teachers engage students so that they can understand and acquire what is being taught. Activities are logical and sequential, with a beginning, a middle, and an end. This means that teachers have a cognitive command of the subject being taught and monitor their performance to ensure a coherent presentation of lessons (Alkin, Doby, & Lindheim, 1990; Brophy & Good, 1986; Stevenson & Stigler, 1992). As previously presented, quality of instructional delivery also includes teachers' expectations for the enacted curriculum (i.e., cognitive demands) and instructional resources such as access to textbooks, calculators, and computers (Boscardin, Aguirre-Muñoz, Chinen, & Leon, 2004; Herman et al., 2000; Porter, 2002; Wang, 1998). Quality of instructional delivery is often measured using direct observations.

OTL Conceptual Framework for This Study

This study used the TIMSS data to examine OTL variations in mathematics in Japanese 8th grade classrooms in the contexts of *yutori* reform, between 1999 and 2007. It also investigated how that potential changes in OTL levels influenced students' measured mathematics achievement levels on the TIMSS tests. As previously mentioned, the public in Japan were concerned that cuts in instructional time and curriculum content in core subjects (i.e., mathematics, Japanese, science) would lead to students learning less than before (Cave, 2003). In other words, students were expected to have less opportunity to learn because of the reforms, with lowered academic standards and poorer outcomes. The OTL conceptual framework used in this study combines the TIMSS conceptual framework of educational opportunity (see Figure 2) with Stevens' multidimensional OTL framework. The current TIMSS tripartite framework mainly addresses the content coverage aspect of OTL. This study sought to further unpack the OTL dimensions manifested in implemented curriculum, drawing on Stevens' multidimensional framework in classrooms.

Specifically, the study utilized the TIMSS teacher survey to identify classroom-level OTL items that fit four variables in the Stevens' framework:

- Content Coverage—The core curriculum topics covered specific to a particular grade level or subject area.
- Content Exposure—The amount of time teachers allocated to covering the content and the depth of the teaching provided.
- Content Emphasis—The emphasis given to certain topics that are part of the core curriculum.
- Quality of Instructional Delivery—The coherence and effectiveness of classroom teaching practices.

The above classroom-level OTL variables are situated within the TIMSS' larger, systems-based framework for examining OTL in a given nation, as shown in Figure 3. As illustrated, this conceptual framework helped formulate the research questions which guided this convergent parallel, mixed-methods analyses.

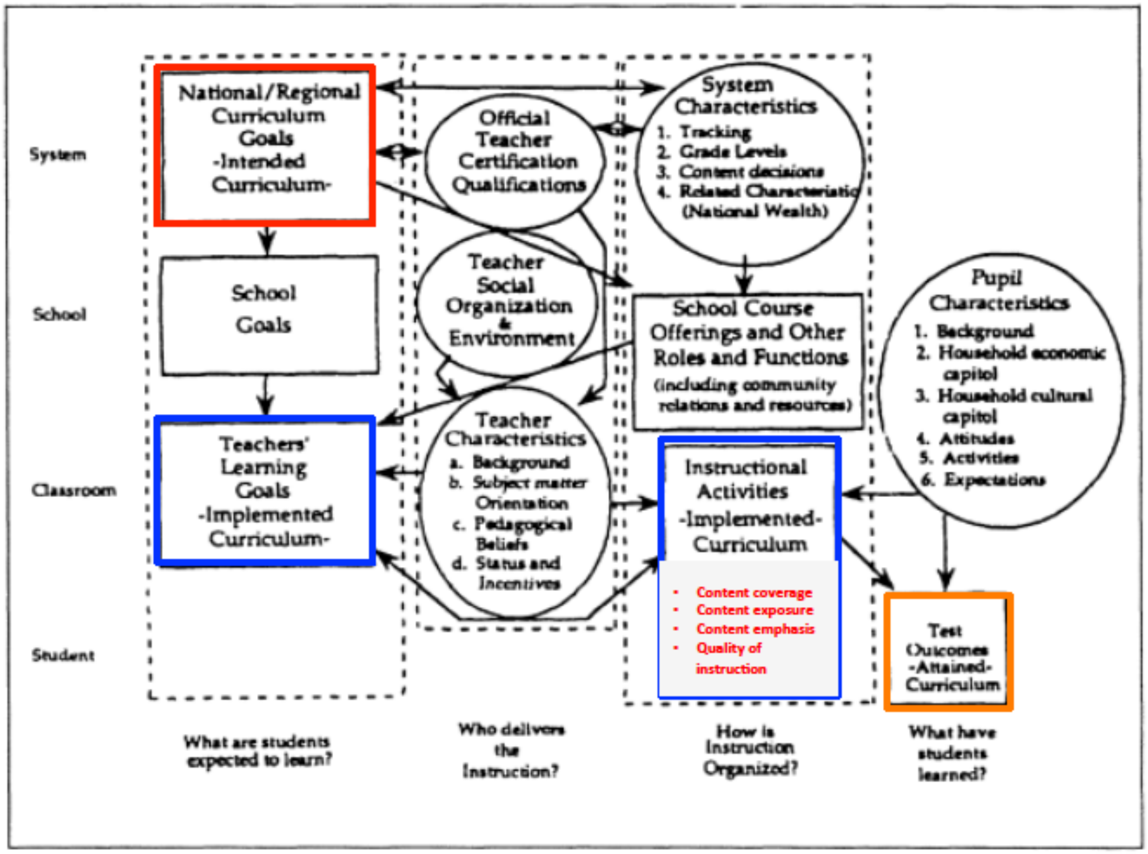

System
School
Classroom
Student
National/Regional Curriculum Goals -Intended Curriculum-
School Goals
Teachers' Learning Goals -Implemented Curriculum-
Official Teacher Certification Qualifications
Teacher Social Organization & Environment
Teacher Characteristics
a. Background
b. Subject matter Orientation
c. Pedagogical Beliefs
d. Status and Incentives
System Characteristics
1. Tracking
2. Grade Levels
3. Content decisions
4. Related Characteristic (National Wealth)
School Course Offerings and Other Roles and Functions (including community relations and resources)
Instructional Activities -Implemented- Curriculum
Content coverage
Content exposure
Content emphasis
Quality of instruction
Pupil Characteristics
1. Background
2. Household economic capitol
3. Household cultural capitol
4. Attitudes
5. Activities
6. Expectations
Test Outcomes -Attained- Curriculum
What are students expected to learn?
Who delivers the Instruction?
How is Instruction Organized?
What have students learned?

Figure 3. OTL conceptual framework of the study
Red fonts indicate the incorporation of the four variables outlined in
Stevens' multidimensional OTL framework

Relationship Between OTL and Achievement

Several meta-analytic studies have examined the effects of school-level OTL measures on achievement (Hattie, 2009; Marzano, 2003; Scheerens & Bosker, 1997; Scheerens et al., 2007). These meta-analysis reviews have unanimously shown OTL to be a positive correlate of achievement according to the Cohen's *d* effect sizes reported: 0.18 (Scheerens & Bosker, 1997), 0.30 (Scheerens et al., 2007), 0.39 (Hattie, 2009), and 0.88 (Marzano, 2003). However, several remarks about these results should be noted. Marzano

(2003) and Scheerens et al. (2007) both included citations from Scheerens and Bosker (1997). Hattie’s (2009) meta-analysis results were based on gifted children.

As a classroom-level independent variable, OTL has also been consistently shown to be a positive correlate of academic achievement (Floden, 2002; Gamoran et al., 1997; Kurz, Elliott, Wehby, & Smithson, 2009; Lafontaine, Baye, Vieluf, & Monseur, 2015; McDonnell, 1995; Schmidt & Maier, 2009; Wang, 1998). The following section presents research that has investigated the link between OTL and student achievement through the three OTL research strands (i.e., studies of instructional time, studies of content overlap, and studies of quality of instruction) and Stevens’ multidimensional OTL framework.

Studies of Instructional Time

This line of OTL studies, namely the work of Carroll (1963), Wiley and Harnischfeger (1974), and Berliner and colleagues (1978, 1981, 1990), examined the relationship between instructional time and achievement. Studies of allocated instructional time have focused mainly on overall instructional time for a subject matter or across subject matters—not on a specific topic area. In a study among 40 Detroit schools, Wiley and Harnischfeger (1974) examined the relationship between the overall instructional hours in an academic year and student achievement. They found statistically significant, positive associations between time and achievement in different academic subjects. Particularly, they found that students, in schools that provided more overall instructional time, had on average reading comprehension and mathematics scores that were 66% and 33% higher than those of students in the control schools. In a research synthesis on teaching, Walberg (1986) reviewed 31 studies that examined the relationship between amount of instructional time and student achievement. His study found a moderate correlation (r = 0.35) after controlling for a host of student-level demographic variables. In a meta-analysis of 21 studies, Scheerens and Bosker (1997) examined the effect of allocated time on student achievement and found a medium effect size for time (Cohen's *d* = 0.39). However, equivocal findings have been common in studies of total instructional time, usually with correlations ranging from zero to moderately positive between time and achievement (Desimone, Smith, & Phillips, 2013; Gamoran, 1987; Karweit, 1985; Oketch, Mutisya, Sagwe, Musyoka, & Ngware, 2012; Walberg & Frederick, 1982).

Berliner and colleagues (1978, 1981, 1990) addressed this gap in their BTES research by defining OTL as the amount of time a student spends engaged with materials and activities, in which a high level of success is attained and the materials and activities are related to outcomes that are valued. The BTES researchers compared the amount of allocated time (i.e., time dedicated for instruction) and engaged time (i.e., time actually spent in academic tasks) in 2nd- and 5th-grade classroom over a 6-year period (Fisher et al., 1981). They found statistically significant correlations between time allocated on a topic and student achievement. In other words, if students spend more time working on a topic, they will learn more about that particular topic (Berliner, 1990; Fisher et al., 1981).

The findings from the BTES were later corroborated by the work of Baker, Fabrega, Galindo, and Mishook (2004) in a review of TIMSS, the IEA Study of Civics Education, and the PISA. Baker et al. (2004) did not find a significant relationship between overall mean instructional time and country-level mean achievement, but they found positive associations when allocated time, specific to the subject tested, was used to predict achievement. Similarly, Hill, Rowan, and Ball (2005) reported that third graders exposed to longer mathematics lessons demonstrated higher mathematics achievement gains. The results from the BTES (Baker et al., 2004; Hill et al., 2005) suggested that overall instructional time is an “often inadequate measure of OTL if it does not specify time spent teaching specific content” (Schmidt & Maier, 2009, p. 554).

Studies of Content Overlap

Another line of OTL work defined OTL in relation to the content covered during instruction, most notably the IEA studies. To date, IEA has conducted six international comparative studies of student achievement, the results of which have supported students’ opportunity to learn the assessed curriculum as a significant predictor of systematic differences in their performance (Schmidt & Maier, 2009). The FIMS showed that OTL was a significant positive correlate of mathematics achievement (Husén, 1967). Based on the pooled data from all countries, correlations between mathematics achievement and OTL ranged from 0.16 to 0.30, depending on the student population (i.e., age or grade) analyzed (Schmidt & Maier, 2009). There was a considerable amount of between-country variance in mathematics scores, which could be attributed to the differences between students’ opportunities to learn the material that was tested (Husén, 1967). While the relationship between OTL and mathematics achievement was modest within countries, the same relationship was substantial between countries. Between-country correlation ranged from 0.40 to 0.80 across the four sampled populations.

Even with a more sophisticated measure of OTL, SIMS results showed a small to non-existent relationship between OTL and achievement within countries (Schmidt & Maier, 2009). Even after accounting for students, parental, teacher, and school characteristics, OTL was only significantly correlated to achievement in one country per topic area (i.e., arithmetic, algebra, geometry, measurement, and statistics) among all participating countries (Schmidt & Kifer, 1989). When looking at the achievement gains on arithmetic, algebra, and geometry subtests, SIMS researchers found OTL to be positively associated to achievement gains among 8th graders in two (of eight) countries after holding student, parental, teacher, and school characteristics constant (Schmidt & Burstein, 1993). According Schmidt and Burstein (1993), the lack of a significant relationship between OTL and achievement could be due to curricular homogeneity such that systems with little within-country variation in OTL demonstrated weaker correlations.

SIMS studies also found a between-country correlation between OTL and mathematics achievement gain of 0.57, which was within the range of correlations found in FIMS (Kifer & Burstein, 1992). Correspondingly, countries whose mathematics curriculum largely focused on one of the specific topics tested tended to demonstrate higher achievement gains on such topics (Schmidt et al., 1992). In other words, "providing more content to more students produces more gain" (Kifer & Burstein, 1992, p. 337).

The findings from the TIMSS were in line with those from FIMS and SIMS: a strong correlation between OTL and mathematics achievement among the 8th graders was found across countries. From extensive curricular data collected from country standards, textbooks, and teachers, the TIMSS reported large differences in the intended and implemented curricula across countries (Schmidt et al., 1999). These differences were found to be significantly correlated to achievement gains (Schmidt et al., 2001).

William Schmidt and his colleagues (2001) conducted a cross-country investigation of achievement gain and five measures of OTL (i.e., national standards, national average instructional time for a topic, proportion of a country's teachers covering a topic, proportion of textbook space devoted to a topic, proportion of textbook space devoted to a topic with greater complexity of student performance expected) which were derived from 20 subscales from 8th grade mathematics in the TIMSS. Statistically significant associations were found between achievement gain and at least one OTL measure in 12 of the 20 mathematics subscales. Most notably, teacher coverage of a topic was found to be strongly correlated with six of the 12 subscales. Schmidt and colleagues (2001) further concluded:

> Curriculum was related to learning in mathematics across countries in seventeen of the twenty tested topic areas as measured by the TIMSS Population 2 test. Further, the relationships with gain involved different aspects of the curriculum other than the content standards measure were represented. (p. 324)

Using the same five OTL measures, Schmidt and colleagues also examined the relationship between OTL and achievement gain within countries. They found statistically significant relationships in 24 of the 29 countries, with estimated R2 values ranging from 0 to 0.67 (Schmidt et al., 2001). Moreover, multilevel analyses were performed on each of the 20 mathematics subscales to examine the OTL-achievement gain relationship while controlling for socioeconomic status and prior achievement. “On 18 of the scales measures of instructional time in key content area were found to be statistically significant to achievement gain. The R2 values ranged from 0.38 to 0.63” (Schmidt & Maier, 2009, p. 553). Lastly, Schmidt et al. fitted within-country causal models to relate content coverage—as defined by the aforementioned five OTL measures—to each other and then to achievement gain. They found that in 19 of the 29 countries, achievement gain in mathematics at the 8th grade was driven statistically by content coverage as defined by the textbook. The estimated R^2 values for the textbook-achievement gain associations ranged from 0.23 to 0.70, where the largest R^2 value came from the model fitted to the Japanese data (Schmidt et al., 2001). This content overlap operationalization of OTL was prominent in several other research studies in the 1970s and 1980s (Borg, 1980; Mehrens & Phillips, 1986; Walker & Schaffarzick, 1974; Winfield, 1987). For their meta-analysis, Scheerens and Bosker (1997) reported an average Cohen's *d* effect size of 0.18 in the 19 content-overlap focused studies they reviewed.

While the foundation for studying OTL internationally is rooted in the TIMSS, the PISA 2012 offered a way to examine the relationship between OTL indicators and students' mathematics literacy. Schmidt et al. (2013) explored the connection between student performance on the PISA 2012 and three student-level OTL indices: OTL related to formal mathematics, OTL related to applied mathematics, and OTL related to word problems. Across the 34 participating countries, they found that OTL related to applied mathematics was statistically significantly related to mathematics literacy and the relationship was quadratic. This implies that more frequent exposure to OTL related to applied mathematics does not add to mathematics performance beyond a certain point.

For all OECD countries, all three OTL indices were found to be statistically significant at the student level (Schmidt et al., 2013). OTL related to formal mathematics and OTL related to word problems were significantly related to performance at the student level as well as the school level in all OECD countries. Interestingly, the OTL related to the applied mathematics-achievement relationship was quadratic in most countries at the student level and the school level. However, the OTL constructs covered in PISA 2012 were confounded by the lack of course or classroom information, such as students' course-taking history and their teachers.

Carnoy and his colleagues (2016) conducted a study that administered the mathematics portion of the PISA 2012 to all 9th grade students who had taken the TIMSS 2011 assessment in Russia. Their unique study included a national sample of Russian students. More than 90% of the 4,893 8th graders from 231 intact classrooms who participated in the TIMSS 2011 took the PISA 2012 survey. Using this longitudinal data set, Carnoy and colleagues were able to estimate the effects of classroom variables on the students' PISA performance. Their results were in agreement with Schmidt et al.'s (2013) multi-nation study on the PISA OTL. Carnoy et al. found that OTL related to formal mathematics had positive, significant effects on students' PISA mathematics performance, even after controlling for students' performance on the TIMSS 2011. They also found that this large positive OTL-achievement relationship was statistically significant for students with low and middle academic resources, but not significant for students with high academic resources. However, Carnoy et al.'s estimates showed a much smaller OTL effect on the PISA scores than did Schmidt and OECD's estimates.

The Content Determinant Project initiated by Porter and his colleagues (1988) was another line of content overlap research focused on students' OTL important content objective. Porter's research focused on examining the content of instruction along two aspects: topics and categories of cognitive demand (Porter, 2002; Porter & Smithson, 2001). In 1997, Gamoran and colleagues compared several approaches of OTL measures in a study of mathematics courses taken among predominantly low-income, low-achieving 10th graders to look for the representation that would have the largest correlation with student achievement gain. They used teacher questionnaires to gather information on the content coverage; they asked teachers both about which mathematics topics were taught and what sort of cognitive demand the instruction asked of the students (Gamoran et al., 1997). Topics and cognitive demand together were labeled as “content coverage” by Gamoran et al. (1997). They found that the highest correlation with achievement gain came when the analysis used a combination of topics and cognitive demand, rather than looking only at topic or demand alone (Gamoran et al., 1997). The correlations were 0.45 with class gains and 0.26 with students when the combination of topics and cognitive demand were used. However, associations with achievement gains were -0.21 at the class level and 0.10 at the student level for topics only, and the same associations were 0.11 at the class level and 0.07 at the student level for cognitive demand only (Gamoran et al., 1997). Furthermore, the highest associations came when content emphasis was distributed in a pattern similar to the distribution of content on the achievement test. This led Gamoran and his colleagues to conclude that content coverage as defined by topics and cognitive demand seems likely to result in student achievement gains (1997).

Following Gamoran and colleagues’ methodology, Smithson and Collares (2007) examined how instructional alignment—as indicated by topics and cognitive demand—with state benchmarks predicted achievement in English Language Arts (ELA) assessment among low-income, low-achieving elementary students in Ohio. Smithson and Collares (2007) found that alignment was positively correlated with achievement gains in ELA ($r = 0.34$, $p < 0.01$). More importantly, alignment accounted for about 71% of the gain in scaled scores.

Taken collectively, the above studies support an empirical association between instructional content and student achievement. In other words, students who have more opportunities to learn the content embedded in tests and/or curricular standards at the appropriate intensity level tend to exhibit achievement gains. However, most of these findings were not generalizable to the average U.S. student population as they were conducted among low-income, low-achieving students. Furthermore, the quality of achievement measures used across these studies varied in quality. Little information is available on the reliability and validity of achievement tests and the tests' alignment with the intended curriculum (Kurz, 2011).

Studies of Quality of Instruction

Quality of instruction research highlights the third strand of OTL research. The operationalization of quality of instruction measures has resulted in a large set of factors related to student achievement. In his research synthesis on teaching, Walberg (1986) reviewed 91 studies that examined the effect of quality indicators on student achievement. His study found strong positive effect for reinforcement and corrective feedback. Saxe, Gearhart, and Seltzer (1999) studied the impact of instructional strategies on students' mathematics achievement by conducting classroom observations and collecting pre- and post-instruction achievement scores. Saxe et al. found that differences in achievement could be attributed to differences in instruction. A wide range of instructional quality variables are available, highlighting the importance for researchers to provide a theoretical and empirical rationale for their particular operationalization of instructional quality.

Studies of Stevens' Multidimensional OTL Framework

Numerous quality of instruction OTL research have followed Stevens' framework to operationalize the OTL variables (Abedi, Courtney, Leon, Kao, & Azzam, 2006;

Boscardin et al., 2004; Herman & Abedi, 2004; Mo, 2008; Wang, 1998). However, none of these studies used all four OTL variables, as outlined in Stevens' framework.

Abedi et al. (2006) examined the relationship between three class-level components of OTL (i.e., student report of content coverage, teacher content knowledge, and class prior mathematics ability) and mathematics achievement tests. Their results indicated that high levels of content coverage, class prior ability, and teacher content knowledge were associated with improved mathematics performance after controlling for individual students' prior mathematics ability.

Boscardin et al. (2004) examined the relationship between OTL and student performance in English and algebra using five OTL measures: teaching experience, teacher expertise in content topics, topic coverage, classroom activities, and assessment strategies and preparations. Boscardin and colleagues demonstrated that OTL operationalized as content coverage was positively correlated with student performance in English and algebra, even when teacher expertise and class-level free/reduced lunch status were controlled. Their results implied that "what and how much a teacher teaches in a class can make a difference in students' performance, regardless of student background" (p. 323).

Herman and Abedi (2004) studied the English language learners' opportunity to learn Algebra I. They operationally defined OTL as content coverage through asking teachers to indicate the content areas covered in their 2-year algebra course and also asking 8th grade students to indicate the content areas their class had covered. They found a strong relationship between the classroom-level OTL measure and student algebra performance, even when controlling for prior mathematics ability (r = 0.72,

$p < 0.01$ for student-reported OTL at the classroom-level; $r = 0.53$, $p < 0.01$ for teacher-reported OTL at the classroom level). More importantly, after accounting for the classroom-level OTL measure, the student-level OTL variable had no significant effect on algebra outcome scores. Their findings supported the use of the classroom-level OTL measure over the student-level one.

Mo et al.'s (2008) study used 2003 TIMSS data to examine the effects of OTL on students' engagement in science and subsequently on science achievement among 8th graders in the United States. Mo et al. used the TIMSS teacher survey items on teachers' instructional activities as a quality of instruction OTL variable. Their results showed that OTL had a significant positive effect on students' engagement in science classroom activities which, in turn, had a significant indirect effect on science achievement.

As one of the first multilevel OTL studies, Wang (1998) found that content coverage, content exposure, and quality of instruction were significant positive correlates of student achievement in science, even after controlling for prior knowledge, gender, and race. Wang further noted that quality of instruction accounted for the largest percentage of variance in hands-on science test scores. Although Wang's study was guided by Stevens' framework, she did not include time on instruction and used an unusual measure of content coverage (i.e., the teachers' predicted pass rate for students on each test item). Despite the shortcomings, Wang showed that quality of instruction can serve as a significant contributor to student achievement, even with other OTL variables in the model.

In summary, this body of literature suggested that OTL is a multidimensional measure, which would be best addressed using a multidimensional OTL conceptual framework like that of Stevens'. Unfortunately, no studies have yet included all four OTL elements highlighted in Stevens' multidimensional framework to examine the relationship between OTL and student achievement.

Chapter III

METHODS

This chapter describes the convergent parallel, mixed-methods design that was used to analyze the data and answer the research questions presented in Chapter I. The chapter first provides brief descriptions of the research design, the population, and the sampling design. Then the chapter introduces the methods employed in this dissertation, including the Process Model for validating OTL measures, and the specific quantitative and qualitative tools used to answer each research question.

Research Design

This study employed a four-phase convergent parallel, mixed-methods design to examine the effects of *yutori* reforms at three levels: *the intended curriculum, the implemented curriculum,* and *the attained curriculum*.

To investigate the *intended curriculum* under *yutori,* I examined and analyzed annual white papers published by MEXT between 1999 and 2011 to map out key components of *yutori* reforms and their implementation timeline, as intended by the Japanese government. I further analyzed descriptive data from the TIMSS NRC Questionnaire to depict the range of topics intended to be covered in the 8th grade mathematics curriculum in Japan by year (TIMSS, 1999, 2003, 2007).

To investigate the *intended curriculum* under *yutori,* I conducted two semi-structured interviews with two researchers who have studied Japanese education for more than 10 years and have done observational research in elementary or junior high schools during *yutori* reform periods. I then derived OTL construct measures drawing on Stevens' OTL framework from the TIMSS Teacher Questionnaire, and validated the OTL construct measures using an iterative process (Chatterji, 2003; in press). Lastly, I investigated the effects of *yutori* reforms on Japanese 8th grade teachers' self-reports of validated OTL measures in mathematics classrooms between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori* cohort).

To investigate the *attained curriculum* under *yutori,* I examined effects of *yutori* reforms on the OTL and mathematics achievement relationship in three stages: (a) I compared the OTL-mathematics achievement relationship between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori cohort*); (b) I examined the OTL effects on students' mathematics achievement within cohorts; and (c) I investigated the SES moderation effect on the relationship between students' mathematics achievement and OTL within cohorts. Table 4 in Chapter I presented an overview of research questions aligned with data sources and analytic methods.

Population and Sampling Design

This study included a secondary analysis of the TIMSS data from Japan collected during the 1999, 2003, and 2007 administrations in selected 8th grade classrooms. The TIMSS has been administered every 4 years since 1995 by the TIMSS and PIRLS International Study Center at Boston College, under the auspices of the International

Association for the Evaluation of Educational Achievement (IEA). The targeted population at 8th grade had the same definition in 1999 and 2003 as follows: "All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing" (Foy & Joncas, 2000, p. 30; Foy & Joncas, 2004, p. 110). In 2007, the targeted population was redefined as the grade that represented 8 years of schooling, counting from the first year of primary or elementary schooling (Joncas, 2008).

Some schools and students within schools were excluded from the national defined target population. In 1999, 2003, and 2007, special needs schools were excluded. In 2007, there were additional within-school exclusions which consisted of classes within general schools with multi-grade organizations, and classes within general schools for disabled children (Joncas, 2008). In these three rounds of TIMSS assessments, Japan had less than 10% of excluded schools and non-participating schools (1999 = 7%, 2003 = 3%, 2007 = 3%, respectively). Tables 5 and 6 provide the basic descriptions of the TIMSS 1999, 2003, and 2007 samples in Japan.

Table 5

Sample Characteristics of Pre-Yutori, Mid-Yutori, and Post-Yutori Cohorts

Background Variable	Pre-*yutori Cohort* (TIMSS 1999)	Mid-*yutori Cohort* (TIMSS 2003)	Post-*yutori Cohort* (TIMSS 2007)
Years of Formal Schooling	8	8	8
Average Age of Students Tested	14.4	14.4	14.4
Total Number of Schools Participated	140	146	146
Total Number of Students Assessed	4745	4856	4312

Note: This table presents the descriptive statistics for the unweighted sample.

Table 6

Characteristics of 8th Grade Teachers Participating in TIMSS 1999, 2003, and 2007

	TIMSS 1999 ($N_{\text{Data Set 1}}$ = 144)		TIMSS 2003 ($N_{\text{Data Set 2}}$ = 146)		TIMSS 2007 ($N_{\text{Data Set 3}}$ = 215)	
Variables	Frequencies	%	Frequencies	%	Frequencies	%
Age						
Under 25	6	4	2	1	11	5
25-29	24	17	17	12	34	16
30-39	57	40	52	36	65	30
40-49	47	33	51	35	76	35
50-59	9	6	22	15	26	12
60 or more	1	1	2	1	3	1
Gender						
Male	104	72	101	69	125	58
Female	40	28	45	31	90	42
Years of Teaching Experience	14 (8)		17 (9)		15 (9)	

Note: This table presents the mean years of teaching experience and its standard deviation in parentheses. Percentages may not add up to 100% due to rounding.

The TIMSS uses a two-tier stratified cluster sampling procedure. In the first stage, schools were selected from the list of all schools with students who fit the defined criteria using probability proportional to size sampling (PPS) techniques. In the second stage, one intact mathematics classroom was randomly selected from each school. These sampling procedures were used during all three TIMSS administrations in Japan, with a caveat. While TIMSS 1999 and 2003 sampled one mathematics classroom per school, TIMSS 2007 sampled "two intact mathematics classrooms per school with more than 230 students, and one classroom otherwise" (Joncas, 2008, p. 390).

Because the TIMSS samples intact mathematics classrooms, the participating students are not independent observations but are, in fact, in random clusters. Students from the same classroom share more common characteristics than students randomly drawn from the whole population of 8th graders in Japan.

Along with mathematics and science assessments, the TIMSS collects background information about students, their teachers, their schools, as well as the mathematics and science curriculum required by the education system in their country. I drew on these surveys extensively to construct OTL measures and covariates that are known to affect OTL and student achievement. The next section describes the methods and the variables used to answer each research question.

Research Question One: Document Analysis of Policy Documents

Research Question One asked: Based on a content analysis of white papers published by MEXT between 1999 and 2011, (a) what was the *intended curriculum* as given by MEXT and (b) why did the motivations and intentions of MEXT change through the pre-*yutori*, mid-*yutori*, and post-*yutori* time periods? To answer this question, I performed document analysis, which entails finding, selecting, appraising, and synthesizing data contained in documents to elicit meaning and gain understanding (Bowen, 2009). Document analysis organizes data into major themes and categories through content analysis (Labuschagne, 2003).

Data Collection Procedures

According to Guest, Namey, and Mitchell (2013), the key to document analysis is finding data sources that are most relevant to the research objectives. Guest et al. suggested asking three guiding questions to narrow the data sources for analysis:

1. What documents, or artifacts have been produced by the current study population that are conceptually related to the research question(s)?

2. What public documents or artifacts contain information that can inform the research question(s)?
3. How accessible are these sources of data?

Using these three guiding questions, I selected documents that were available in online, open-access archival repositories of MEXT (http://www.mext.go.jp/en/publication/whitepaper/index.htm). These documents were annual white papers published by MEXT that introduced new and important policies in the areas of education, science and technology, sports, and culture for their year of publication. The information provided in these white papers best represents the Japanese Government's intentions and actions for *yutori* reforms.

Data Analysis

Document analysis involves an iterative process of skimming, reading, and interpreting (Bowen, 2009). Following the methodology suggested by Corbin and Strauss (2008), I began documentary analysis by doing a "first-pass" document review to identify *yutori*-related information while taking notes on the background and context of *yutori* reform. I then summarized those notes in a chronological table that showed the progression of *yutori* reform from 1999 to 2011. According to Yin (2014), the chronological approach is most appropriate when events unfold and follow a process.

In my re-reading of the white papers, I used sections from the white papers as units of analysis and coded the data for events related to *yutori* reforms. Coded events and their description were then added to the chronological table. Then I organized the coded data into categories and identified themes emerging from the data. The chronological table was updated one last time with emergent themes.

To better illustrate the inter-relations I saw between time points and specific events, I converted the chronological table into a concept map. Concept maps are graphical tools for organizing and representing the relationships between concepts or events (Novak & Cañas, 2008).

Triangulation of Themes

According to Creswell and Plano-Clark (2018), triangulation is a procedure whereby inquiries can be enhanced by corroborating evidence from different and multiple sources to provide better perspectives on the topic of study. Triangulation or multiple methods of data collection also strengthen reliability as well as internal validity (Merriam, 2009; Patton, 2002). To cross-validate the themes derived from the document analysis, I compared and contrasted the document analysis findings to the intended 8th grade mathematics curriculum, as given by the TIMSS national research coordinator surveys. The compilation of the intended 8th grade mathematics curriculum is described in the methods used for Research Question Two.

Research Question Two: Descriptive Analysis of TIMSS Curriculum Questionnaires

Research Question Two asked: Based on the perceptions of National Research Coordinators as reported on the TIMSS surveys in 1999, 2003, and 2007, to what extent did the intended national mathematics curriculum change in terms of content coverage as *yutori* reforms were implemented between 2003-2007 in Japan?

National Research Coordinators (NRC) were officials who oversaw educational policies and practices at the national or regional level. They were asked to indicate the array of TIMSS topics included in their country's intended curriculum at the 4th and 8th grades on the TIMSS Curriculum Questionnaire.

Data Collection Procedures

I downloaded TIMSS International Mathematics Reports for 1999, 2003, and 2007 from the TIMSS and PIRLS International Study Center on the Boston College website (https://timssandpirls.bc.edu).

Data Analysis

Quantitative descriptive statistical analyses were performed to answer this research question. Descriptive statistical analysis provides data summaries about the sample and construct measures to describe the basic features of the data in a study.

First, I reviewed TIMSS International Mathematics Reports for 1999, 2003, and 2007 and extracted the intended 8th grade TIMSS mathematics topics, as provided by the NRC. Since the information about content coverage was already summarized in percentages in the TIMSS International Mathematics Reports, I did not perform any additional analysis. I simply merged the three rounds of descriptive data to compare and contrast the 8th grade mathematics curriculum intended by MEXT.

Research Question Three: Researcher Interview Data

Research Question Three asked: From two school-based researchers' observations of Japanese junior high schools during the *yutori* reform period: (a) what roles did teachers and schools play in implementing *yutori* reform directives, (b) what support and barriers did schools and teachers face as they attempted to follow *yutori* reform guidelines, and (c) what were other contextual factors affecting the implementation of *yutori* reforms in Japanese schools and classrooms?

Data Collection Procedures

I conducted semi-structured interviews (see Appendix A for interview guides) with two researchers who had studied *yutori* reforms extensively. The interviewee sample was selected using expert sampling, which is a type of purposive sampling technique that is used to glean knowledge from individuals with particular expertise (Lavrakas, 2008).

Using Google Scholar, I used keyword "*yutori*" to search for researchers who have written about *yutori* reforms within the last 10 years. Of those researchers, I sent invitational emails to the four most well-published researchers on February 15, 2018 (see Appendix B for email invitation). I recruited two researchers by February 19, 2018. I interviewed the first researcher in person on February 28, 2018 and the second researcher via Skype on March 1, 2018.

During both interviews, I took notes on what I observed and heard. I hired a transcriber to transcribe both interviews. After transcription, I listened to each interview and checked for accuracy.

Data Analysis

The analysis of interview data was based on an inductive approach geared to identify patterns in the data through thematic codes. Inductive analysis allows analysis of patterns, themes, and dimensions that emerge from the data without imposing any hypotheses prior to data collection (Patton, 2002). I reviewed the transcripts and my notes line-by-line and identified themes in the data (Bogdan & Biklen, 1998). Next, I considered the emergent themes in light of my research questions and regrouped the themes into four categories:

1. Define (How were the *yutori* reforms implemented in schools and classrooms?)
2. Support (What support did the schools and teachers receive as they attempted to implement *yutori* reforms?)
3. Barriers (What barriers did the schools and teachers encounter as they attempted to implement *yutori* reforms?)
4. Contexts (What other contextual factors affected students' mathematics OTL in classrooms and at schools?)

Triangulation of Themes

Triangulation of multiple research methods was used again as a way of enhancing internal validity (Creswell, 2014; Merriam, 2009). To cross-validate the emergent themes from the interview analysis, I compared and contrasted the interview analysis findings to the implemented 8th grade mathematics curriculum, as given by the TIMSS teacher surveys. The compilation of the implemented 8th grade mathematics curriculum is described in the methods used to address Research Question Five.

Research Question Four: Validation of TIMSS Construct Measures of OTL

Research Question Four asked: To what extent is Stevens' multidimensional framework suggesting four interrelated OTL constructs, upheld in the TIMSS 1999, 2003, and 2007 survey data from samples of participating Japanese 8th grade mathematics teachers? The purpose of this research question was to derive and validate OTL measures using theory in order to examine changes in the implementation levels of OTL at the pre-, mid-, and post-*yutori* reform periods.

The Process Model (Chatterji, 2003, in press) is an iterative instrument design and validation methodology that relies on different but relevant kinds of validity evidence to assure the psychometric quality and meaningfulness of construct measures to best meet inferential needs of users. Figure 4 illustrates how the Process Model was applied in this study.

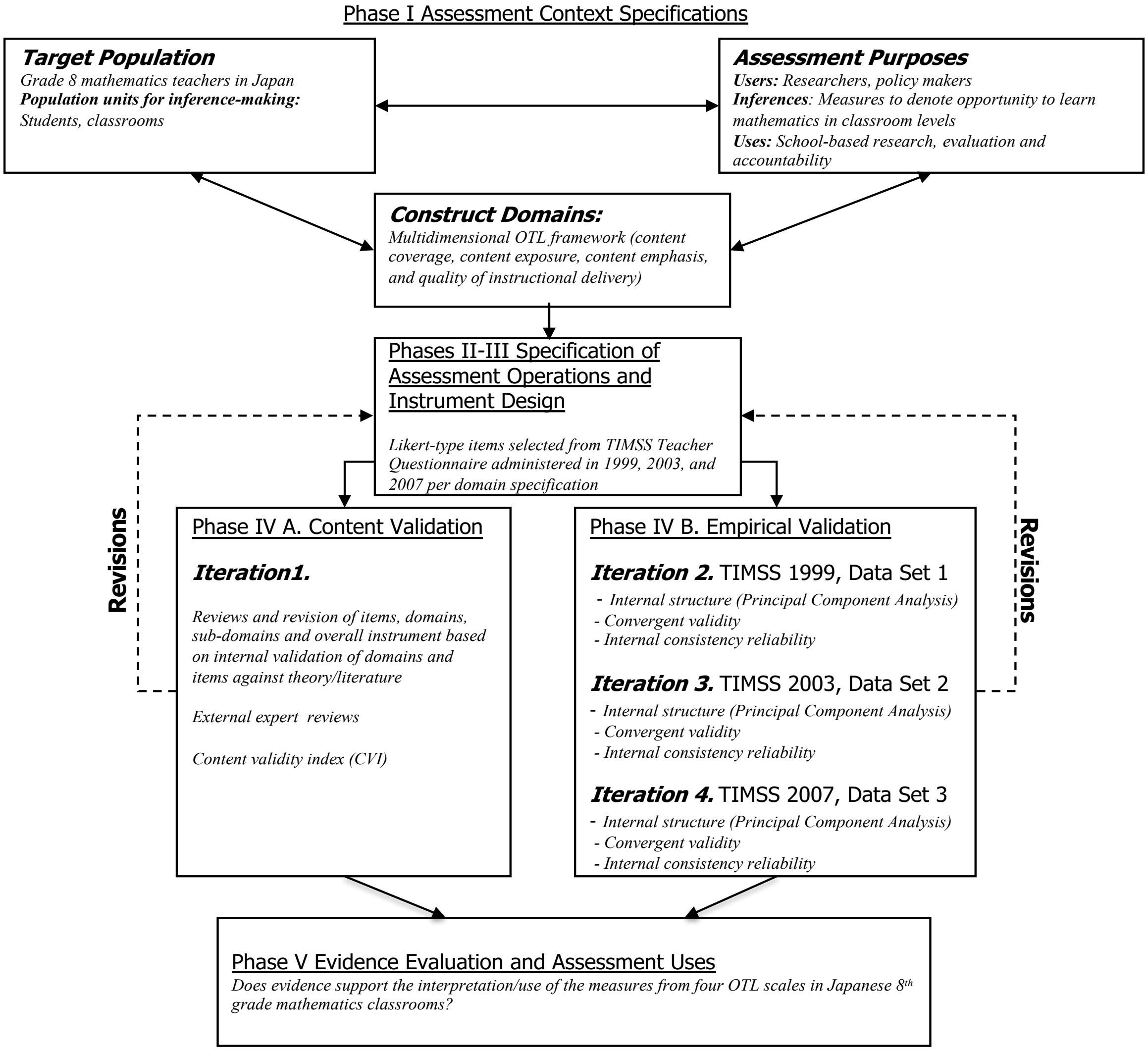

Phase I Assessment Context Specifications
Target Population
Grade 8 mathematics teachers in Japan
Population units for inference-making:
Students, classrooms
Assessment Purposes
Users: Researchers, policy makers
Inferences: Measures to denote opportunity to learn mathematics in classroom levels
Uses: School-based research, evaluation and accountability
Construct Domains:
Multidimensional OTL framework (content coverage, content exposure, content emphasis, and quality of instructional delivery)
Phases II-III Specification of Assessment Operations and Instrument Design
Likert-type items selected from TIMSS Teacher Questionnaire administered in 1999, 2003, and 2007 per domain specification
Revisions
Revisions
Phase IV A. Content Validation
Iteration1.
Reviews and revision of items, domains, sub-domains and overall instrument based on internal validation of domains and items against theory/literature
External expert reviews
Content validity index (CVI)
Phase IV B. Empirical Validation
Iteration 2. TIMSS 1999, Data Set 1
- Internal structure (Principal Component Analysis)
- Convergent validity
- Internal consistency reliability
Iteration 3. TIMSS 2003, Data Set 2
- Internal structure (Principal Component Analysis)
- Convergent validity
- Internal consistency reliability
Iteration 4. TIMSS 2007, Data Set 3
- Internal structure (Principal Component Analysis)
- Convergent validity
- Internal consistency reliability
Phase V Evidence Evaluation and Assessment Uses
Does evidence support the interpretation/use of the measures from four OTL scales in Japanese 8th grade mathematics classrooms?

Figure 4. Validation methodology for uses of OTL measures at classroom level
Adapted from Chatterji (2003)

Phase I: Specifying the Assessment Context

Phase I of the Process Model begins by specifying the context of assessment use. The three questions asked at this stage are: (a) what to measure, which refers to the targeted OTL construct domains and measures to be generated at the classroom level, (b) whom to measure, which refers to population units (teachers) from whom OTL-based inferences will be drawn, and (c) why measure, which refers to the uses to be made from OTL measures derived in classroom, research, or policy contexts. Specifying purposes also involves identifying assessment users, inferential needs based on OTL scale scores, and uses intended by each user (Chatterji, 2003, in press). Beyond this research study, users of the OTL measures could expand to other researchers and policymakers. Specifying population requires identifying the population units from whom inferences would be drawn from the construct measures and their background characteristics (Chatterji, 2003, in press).

Phases II-III: Specifying the Assessment Operations

Phase II of the Process Model involves developing domain specifications with observable indicators of the constructs, grounded in existing theory. I developed a detailed set of assessment specifications for the OTL construct measures based on literature reviews and existing theory (Chatterji, 2003) presented in Chapter II. Common items from the TIMSS 1999, 2003, and 2007 Teacher Questionnaire were reviewed vis-à-vis the specified indicators of the multidimensional OTL framework (Stevens, 1996). This initial review of items resulted in 61 items that reflect a reasonable and substantive match for the domain specification are presented in Table 7.

Phase IV: Content Validation and Empirical Validation

Phase IV of the Process Model entailed a content validation study (Phase IV A) and empirical validation studies (Phase IV B) to examine the internal structure as well as convergent validity and reliability, and to evaluate the overall quality of the proposed OTL construct measures.

Phase IV A. In the content validation phase, I asked two external experts to evaluate the OTL items I selected from the TIMSS inventory matched with the theoretical indicators of the multidimensional OTL framework (Stevens, 1996). Experts in classroom pedagogy in mathematics and measurement were consulted to help make improvements to the four proposed OTL scales. Following each content-based review, refinements were made to the operational definition of OTL to ensure the content relevance and content representativeness of the final scales (Chatterji, 2003). Following the content-based reviews, I estimated the content validity of OTL scales using a content validity index (CVI) proposed by Polit, Beck, and Owen (2007). CVI is a measure of inter-rater agreement used to estimate and quantify content validity (Polit et al., 2007). To capture inter-rater agreement, I calculated the kappa statistic using the CVI value for each item adjusted for chance agreement. Using guidelines suggested by Polit and colleagues (2007), items with kappa of approaching 1.0 were considered as having excellent matches to the proposed indicators, whereas items with kappa approaching 0 were considered as having a poor fit with the proposed indicators.

Phase IV B. The second part of Phase IV involves empirical validation, which includes appropriate forms of "data collection, observation, and analysis of data generated by the instrument devised, followed by evaluation of the psychometric quality of the results" (Chatterji, 2003, p. 110). The present study used three separate data sets—Data Set 1 (TIMSS 1999; $N_{\text{Data Set 1}} = 144$), Data Set 2 (TIMSS 2003; $N_{\text{Data Set 2}} = 146$), and Data Set 3 (TIMSS 2007; $N_{\text{Data Set 3}} = 215$)—prepared using data collected from the TIMSS Teacher Questionnaire.

A Principal Components Analysis (PCA) with varimax rotation was performed to examine evidence of how well the extracted internal structure matched the item composition given by the theoretically-specified and content-validated measures. PCA was selected over other exploratory factor analysis (EFA) techniques because PCA has a distinct advantage "if factor scores are to be used as independent variables or dependent variables in other analyses" (Tabachnick & Fidell, 2007, p. 646). PCA extracts maximum variance from the data set by orthogonal components (Tabachnick & Fidell, 2007). Different from Components were identified based on eigenvalues > 1, observed breaks in the scree plot, and cumulative percent variance explained. Items were not interpreted if (a) a component was defined by fewer than three items, and (b) item sets were not validated across at least two data sets with loadings more than |.32| (Tabachnick & Fidell, 2007). The PCA findings were evaluated with reference to Stevens' OTL framework and the theoretically-derived domain structure and scores (Chatterji, 2003).

Specifically, in this study, PCA was used to evaluate evidence of the internal structure of the theoretically-specified measures for the 9 items from *content exposure* and the 24 items from *quality of instructional delivery*. Items drawn from the *content coverage* and *content emphasis* dimensions of Stevens' OTL framework were not included in the empirical validation phase because those items were descriptive in content (e.g., specific math topics covered) and contained categories that were measured on a nominal scale.

Following PCA, convergent validity was assessed with the Pearson product-moment correlations among construct measures or inter-factor correlations. In this study, convergent validity, as expected by the literature, was estimated through the examination of inter-factor correlations. Per the literature review, the theoretically-specified OTL measures were expected to correlate with each other in the order of .30-.50 (Herman et al., 2000).

Finally, the internal consistency reliability of the scale scores or the validated measures was examined using Cronbach's alpha coefficients, and item analysis with methods from Classical Test Theory were conducted. Internal consistency reliability estimates reveal "the extent to which items from the same domain or subdomain generate consistent patterns of response for individual respondents" (Chatterji, 2003, p. 435). Scales or subscales with Cronbach's alpha coefficient of 0.70 or above are considered acceptable (Crocker & Algina, 2006). Item analysis is useful for selecting the best functioning items, given the stated purposes for the assessment in target populations. Item descriptive and homogeneity statistics such as *item mean*, *item variance*, and *item-to-total score correlation* were described and examined. The *item-to-total score correlation* was estimated using the *point-biserial correlation coefficients* (r_{pb}), which presents the correlations between item response distribution with the total score distribution and the item deleted from the total score calculation (Chatterji, 2003). Negative or low *item-to-total score correlations* ($r_{pb} < 0.2$) suggest poor items that should be either removed or revised before the final instrument is assembled.

Table 7

Domain Specifications and Survey Items by OTL Indicator

Construct	Indicators	Questions	Scale
1.0 Content coverage (Total Items = 20)	1.1 Teacher arranges for all students to have access to the core curriculum 1.2 Teacher arranges for all students to have access to critical subject matter 1.3 Teacher ensures that there is curriculum content and test content overlap	The following list includes the main topics address by the TIMSS math test. Check the response that describe when students in your math class have been taught each topic. ***Major topics: Number, Algebra, Geometry, Measurement, Data and Chance***	1 = Taught before this year 2 = Mostly taught this year 3 = Not yet taught
2.0 Content exposure (Total Items = 9)	2.1 The amount of time teachers allocated to covering the content 2.3 Time devoted to a subject area (i.e. math)	How many minutes per week do you teach math to your math class?	Continuous
	2.2 Time allotted to students to learn 2.3 Time devoted to a subject area (i.e. math)	If you assign math homework, how many minutes of math homework do you usually assign your students?	1 = Less than 15 minutes 2 = 15-30 minutes 3 = 31-60 minutes 4 = 61-90 minutes 5 = More than 90 minutes
	2.1 The amount of time teachers allocated to covering the content 2.2 Time allotted to students to learn	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on each of the following activities? •homework review •lecture-style presentation by teacher •teacher-guided student practice •re-teaching and clarification of content/procedures •student independent practice •tests and quizzes •other	Percentages
3.0 Content emphasis (Total Items = 6)	3.1 Teacher selects topics within the curriculum to teacher (i.e., as a major topic, a minor review or not taught at all) 3.2 Teacher selects the dominant student ability level to teach the curriculum 3.3 Teacher selects which skills and concepts to teach and which to emphasize to all groups of students	What subject matter do you emphasize the MOST in your math class? **Subjects: Number, Algebra, Geometry, Combined Algebra and Geometry, Combined Algebra, Geometry, Number, etc., Other.**	Percentages
		In your view to what extent do the following limit how you teach your math class? •students with different academic abilities •students who come from a wide range of backgrounds •students with special needs •uninterested students •disruptive students	1 = Not at all 2 = A little 3 = Some 4 = A lot

Table 7 (continued)

Construct	Indicators	Questions	Scale
4.0 Quality of instructional delivery (Total Items = 26)	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.2 Teachers uses teaching practices (coherent lessons) to produce students' academic achievement	How often do students in your math class use calculators for the following activities: •checking answers •routine computation •solving complex problems •exploring number concepts	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	How often do you usually assign math homework?	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	If you assign math homework, how often do you assign each of the following tasks? •problem/question sets in textbook •small investigation or gathering data •finding one or more uses of the content covered	1 = Never 2 = Sometimes 3 = Always
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	If you assign math homework, how often do you do the following tasks? •record whether or not the homework was completed •have students correct their own assignments in class •use it as a basis for class discussion •use it to contribute towards students' grades or marks	1 = Never 2 = Sometimes 3 = Always
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.3 Teacher has a cognitive demand of the subject matter	How often do you interact with other teachers to discuss about how to teach a particular concept?	1 = Never 2 = 1-3 times a month 3 = 1-3 times a week 4 = Almost everyday
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	In your math lessons, how often do you usually ask students to do the following: •explain the reasoning behind an idea •represent and analyze relationships using tables, charts, or graphs •work on problems for which there is no immediately obvious method of solution •use computers to solve exercises or problems •write equations to represent relationships •practice computational skills	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson

Table 7 (continued)

Construct	Indicators	Questions	Scale
4.0 Quality of instructional delivery (Total Items = 26)	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.2 Teacher uses teaching practices (coherent lessons) to produce students' academic achievement	Do you use a textbook in teaching math to your class?	1 = No 2 = Yes
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	Are the students in the TIMSS class permitted to use calculators during math lessons?	1 = unrestricted use 2 = restricted use 3 = calculators are not permitted
	4.3 Teacher has a cognitive demand of the subject matter	How well prepared do you feel you are to teach the following topics? •perimeter, area, and volume •coordinate geometry •algebraic representation •solving linear equations and inequalities •simple probabilities-understanding and calculations	1 = Not well prepared 2 = Somewhat prepared 3 = Very well prepared

Phase V: Evidence Evaluation and Assessment Use With Reference to Phase I

Phase V of the Process Model entailed a comprehensive evaluation of all the evidence for the proposed classroom-level inferences and uses intended with OTL measures, in this study and beyond. This phase asks the question: Does the collective evidence obtained through Phase IV support the intended interpretations and uses of the OTL construct measures? The results of the validation stages are provided in Chapter IV, and feed into the next stage of the mixed-methods study.

Research Question Five: Changes in OTL Over Time

Research Question Five asked: Using the validated OTL measures per Stevens' framework, to what extent did the *yutori* curricular reforms affect changes in OTL levels over time in 8th grade mathematics classes, as given by the TIMSS teacher survey data from 1999, 2003, and 2007? To address this question, I examined the effects of OTL on aggregated student achievement at classroom level. This analytic approach was reasonable and critical because OTL is a teacher-level variable, per Stevens' (1996) OTL framework.

Measures

I first merged the three TIMSS datasets and created the measures as described below.

OTL variables were the dependent variable. The OTL variables were created using the derived and validated item sets, as described in prior sections of this paper. To create each OTL variable, I summed and standardized the data from items on the scale. These OTL variables were treated as continuous variables measured on interval scales. Two OTL variables were retained after content validation and empirical validation: *quality of instructional delivery* OTL and *instructional time* OTL.

Year of reforms variable was the independent variable. To compare the mean differences in mathematics achievement over time, I created a Year variable based on the TIMSS assessment cycle that included three categories: (a) TIMSS 1999, (b) TIMSS 2003, and (c) TIMSS 2007.

Descriptive Statistics

To address Research Question Five, I first reviewed data from the TIMSS Teacher Questionnaires to extract information corresponding to the theoretically-derived *content coverage* and *content emphasis* aspects of Stevens' OTL framework. As indicated previously, the *content coverage* and *content emphasis* items yielded descriptive data measured on a nominal scale. As such, I presented the changes in *content coverage* and *content emphasis* over time using descriptive analysis of TIMSS teacher surveys. Topic coverage data for the intended mathematics curriculum were extracted from the 1999, 2003, and 2007 TIMSS International Mathematics Reports (Mullis et al., 2004; Mullis et al., 2008; Mullis, Martin, Gonzalez, Gregory, & Garden, 2000) and descriptive statistics were computed.

One-way Analysis of Variance

To test if there were significant differences in classroom-level OTL measures between the three consecutive rounds of the TIMSS, a one-way analysis of variance (ANOVA) with Tukey HSD post-hoc tests was conducted. The dependent variable in these analyses was the validated OTL measure; the independent variable was Year of TIMSS data collection (i.e. 1999, 2003, and 2007). Separate ANOVA models were run with each validated OTL measure.

The null hypothesis in this case was one of no difference in mean classroom-level OTL measures across the TIMSS 1999, 2003, and 2007. Considering the significant cuts in mathematics instructional time outlined in *yutori* reforms, the null hypothesis was expected to be rejected. One-way ANOVA tests show if different groups of cases have reliable mean differences on a dependent variable (Tabachnick & Fidell, 2007). The one-way ANOVA model assumes normality of sampling distributions, homogeneity of variance, independence of errors, and absence of outliers (Tabachnick & Fidell, 2007).

Research Question Six: Changes in OTL Over Time

Research Question Six asked: To what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement:

(a) at the classroom level, to what extent did the observed changes in OTL levels over time affect changes in aggregated 8th grade students' mathematics achievement between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori*)?

(b) using a multilevel approach, to what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement within cohorts?

(c) did OTL moderate the relationship between students' mathematics achievement and socioeconomic background within cohorts, as measured by the TIMSS student assessments?

Part (a) of this question was addressed using analysis of covariance (ANCOVA); Parts (b) and (c) of this question were addressed using a series of hierarchical linear models (HLM).

ANCOVA Measures

Aggregated Mathematics Achievement Score was the dependent variable for ANCOVA models. I created this measure by aggregating students' mathematics scores on the TIMSS assessment at the class level. This is a continuous variable measured on an interval scale.

OTL variables are the covariates for the ANCOVA models. Two OTL variable were retained after content validation and empirical validation: *quality of instructional delivery* OTL and *instructional time* OTL. As before, the OTL variables were created using the derived and validated items as described in prior sections of this paper, and standardized. These OTL variable were treated as continuous variables measured on an interval scale.

The variable, *Year of reforms,* was the independent variable for the ANCOVA models. To compare the mean differences in mathematics achievement over time, I created a variable called the TIMSS assessment cycle that included three categories: (a) TIMSS 1999, (b) TIMSS 2003, and (c) TIMSS 2007. This variable was used in the ANOVA and ANCOVA models to analyze changes in mathematics achievement levels at the classroom level as *yutori* reforms progressed.

HLM Measures

The Mathematics Achievement Score is the outcome variable for HLM models. It is each student's mathematics score on the TIMSS assessment. This is a continuous variable measured on an interval scale. The TIMSS mathematics achievement scales are Item Response Theory (IRT)-based, ranging from 0-1000, and were established as a part of the TIMSS 1995, based on the participating countries at the time (Gonzalez & Miles, 2001). A mean of 500 and a standard deviation (SD) of 100 were set to reflect the mean and SD of overall achievement across countries in 1995. Due to time constraints and the use of rotated test booklets, students do not answer the same number of items in each specific content area. Therefore, the TIMSS does not produce individual test scores for students; rather, the TIMSS produces a set of five plausible values for each student in mathematics based on aggregated sample statistics. Plausible values are designed to reduce the effect of measurement error in the estimation of population-level parameters (Mislevy, 1991). The TIMSS draws five plausible values at random from the conditional distribution of proficiency scores for each student (Gonzalez & Miles, 2001). Each plausible value provides information about each student's proficiency level as well as information about the uncertainty in the score. In the HLM analysis, the parameter estimates for classroom and students are based on the average parameter estimates from separate HLM analyses of the TIMSS plausible values (Raudenbush, Bryk, & Congdon, 2000).

Two OTL variables in HLM were: *quality of instructional delivery* OTL and *instructional time* OTL. These validated OTL variables are the main classroom-level independent variables in all HLM models. Each OTL scale measure was created by summing the standardized scores of all the validated items that defined the scale. These were treated as continuous variables measured on an interval scale.

Student-level background variables were used as covariates to adjust for potential sources of variance in HLM analysis. The covariates are student's gender (STSEX), student's age (STAGE), and student's socioeconomic status (STSES). STSEX is a dichotomous variable where 1 = female and 0 = male. AGE in years is a continuous variable. I used student's home educational resources (HER) as a proxy for students' SES (Mullis & Martin, 2013). HER is an index variable constructed by IEA based on students' answers to six items: (a) Number of books in the home, (b) Having a study desk for own use, (c) Having a computer, (d) Having a dictionary, (e) Father's education, and (f) Mother's education. HER variable included three response categories (1 = low, 2 = medium, and 3 = high). I standardized the HER variable to create the STSES variable, which has a mean of 0 and an SD of 1.

Teacher-level background variables were used as covariates to adjust for potential sources of variance at the classroom level in HLM analyses. These covariates included teacher's sex (TCHSEX), age (TCHAGE), years of teaching experience (TCHEXP), number of students in their mathematics class (NUMSTU), and aggregated classroom-level SES (CLSES). TCHSEX is a dichotomous variable where 1 = female and 0 = male. TCHAGE is an ordinal variable that is recoded as 1 = under 30 years old, 2 = 30-39 years old, 3 = 40-49 years old, 4 = 50-59 years old, and 5 = 60 years old or more. TCHEXP and NUMSTU are both continuous variables measured on a ratio scale. CLSES is a continuous variable measured on an interval scale. I generated CLSES by summing STSES for all students in a class. Table 8 presents definitions, metrics, and descriptive statistics for all variables included in the analysis.

Analysis of Covariance

I conducted an analysis of covariance (ANCOVA) to determine whether the classroom-level means of student achievement in mathematics were equal across different administrations of TIMSS, while statistically controlling for OTL effects. The dependent variable was the mathematics achievement of 8th graders as the dependent variable. The independent variable was rounds of TIMSS (i.e., 1999, 2003, and 2007) and the covariates were the validated OTL measures. Separate ANCOVA models were conducted for each OTL measure. At the end, I performed one final ANCOVA model with all the OTL measures included.

Hierarchical Linear Models

A series of Hierarchical Linear Models (HLM) were used to examine the association between classroom-level OTL and students' mathematics achievement because of the nested data structure of classrooms. The nested data structure violates the independence assumption required by traditional statistical analyses such as ANOVA and ordinary least-square (OLS) multiple regression (Raudenbush & Bryk, 2002). These

Table 8

Definitions, Metrics, and Descriptive Statistics of Variables in the Study

		TIMSS 1999 ($n_{student}$ = 3152) ($n_{teacher}$ = 94)		TIMSS 2003 ($n_{student}$ = 3780) ($n_{teacher}$ = 116)		TIMSS 2007 ($n_{student}$ = 4296) ($n_{teacher}$ = 159)	
Variable	Definition and description	Mean	SD	Mean	SD	Mean	SD
Dependent variable							
Mathematics achievement score	Student's mathematics scores on the TIMSS (Average of five plausible values)	579.40	78.00	563.18	79.35	572.82	85.34
Student-level background variable							
Female student	0 = male; 1 = female	0.49		0.50		0.50	
Student age	Years	14.38	0.29	14.40	0.32	14.47	0.29
Student SES [a]	Standardized scores of HER	-0.01	0.99	0.00	1.00	-0.02	1.00
Classroom-level background variable							
Female teacher	0=male; 1=female	0.29		0.31		0.43	

Teacher age	1 = under age 25;
	2 = age 25-29;
	3 = age 30-39;
	4 = age 40-49;
	5 = age 50-59;
	6 = over age 60
		3.18	0.94	3.51	0.96	3.41	1.11

Teacher experience	Years	14.16	7.75	16.93	8.61	15.27	9.55
Class SES [a]	Aggregated scores of student SES	-0.02	0.25	-0.01	0.29	-0.05	0.37
Number of students	Number of students in the mathematics class	36.07	3.49	35.00	4.61	31.92	9.17
OTL: Quality of instructional delivery [b]	Aggregated standard scores of validated *quality of instructional delivery* scale	0.13	6.99	0.07	6.46	-0.10	6.31
OTL: Instructional time [c]	Standard score for weekly instructional time given to mathematics classes	-0.06	1.10	0.09	1.07	0.01	0.99

Notes: Descriptive statistics are presented for the unweighted samples.

[a] Number of books at home was used as a proxy for SES for TIMSS 1999 because parental education information was not available.

[b] Mean, SD, and range for the raw HER scale scores are: M = 32.00, SD = 4.99, Range: (21.00-46.00) for TIMSS 1999; M = 33.23, SD = 3.90, Range: (25.00-43.00) for TIMSS 2003; M = 34.91, SD = 3.90, Range: (25.00-45.00) for TIMSS 2007.

[c] Mean, SD, and range for the raw data on the instructional time are: M = 197.97, SD = 17.23, Range: (45.00-300.00) for TIMSS 1999; M = 157.47, SD = 32.04, Range: (75.00-350.00) for TIMSS 2003; M = 156.62, SD = 21.28, Range: (59.00-300.00) for TIMSS 2007.

independence violations tend to inflate Type I errors and yield biased parameter estimates (Raudenbush & Bryk, 2002). HLM provides an integrated strategy for handling the aforementioned problems.

This study endeavored to explain the variations in student outcomes by first decomposing observed relationships into between- and within-classroom components. A series of two-level models, with Level 1 as students (i.e., between individuals and within classrooms) and Level 2 as classroom (i.e., between classrooms), were estimated for each round of TIMSS. Two-level models partition the outcome variance into between- and within-classroom portions, allowing for a more accurate estimation of classroom-level effects on individual-level outcomes (Hox, 2010).

Model 1: Fully unconditional model. The first stage is an unconditional model, which is the simplest HLM model with no predictor variables from any level (Raudenbush & Bryk, 2002). This model is used to estimate how much variance in measured achievement is attributed to the classroom level and the student level. This model partitions total variance in students' mathematics achievement into between-classroom (τ) and within-classroom (σ^2) components. The variance estimates were obtained by fitting an HLM in which each student's mathematics achievement scores (Y_{ij}) were estimated via the classroom mean (β_{0j}) and the unique errors associated with that student (r_{ij}). The classroom mean was estimated by the grand mean (γ_{00}) and the random effects for each classroom (u_{0j}).

$$Y_{ij} = \beta_{0j} + r_{ij} \qquad r_{ij} = N(0, \sigma^2)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} = N(0, \tau_{00})$$

where

Y_{ij} represents the mathematics achievement score (here, plausible value) of each student *i* in classroom *j*;

β_{0j} represents the mean mathematics achievement score for classroom *j,* here, aggregated plausible values);

r_{ij} represents the residual for student *i* in classroom *j*;

$\gamma 00$ represents the grand mean mathematics achievement score across **classrooms** in Japan;

u_{0j} represents the random effect for classroom *j*;

$i = 1, 2, \ldots, n_j$ students in classroom *j*;

$j = 1, 2, \ldots, j$ classroom.

From this model, the interclass correlation (ICC) was estimated by taking the ratio of between classroom variance over the total variance. According to the literature, the ICC must exceed 10% to meet the necessary conditions for performing HLM analysis (Lee, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

Model 2: Random-intercept model with student- and class-level covariates. In Model 2, student and teacher background variables along with class-level SES variable were added to Model 1 to adjust for these potential sources of variance. Model 2 was estimated to determine the amount of unexplained within-classroom variance in mathematics achievement that could be explained by student background characteristics. To yield meaningful interpretation, the student-level SES variable, $(SES)_{ij}$, is group mean centered, and the class-level SES variable, $(\overline{SES})_j$, is grand mean centered.

$$
\begin{aligned}
Y_{ij} = \beta_{0j} + \beta_{1j}\,(\mathrm{STSEX})_{ij} + \beta_{2j}\,[(SES)_{ij} - (\overline{SES})_j] & \qquad\qquad r_{ij} = N\,(0, \sigma^2) \\
+ \beta_{3j}\,(\mathrm{STAGE})_{ij} + r_{ij} & \\
\beta_{0j} = \gamma 00 + \gamma 01[(\overline{SES})_j - (\overline{SES})_{..}] & \\
+ \gamma 02\,(\mathrm{TCHEXP})_j + \gamma 03\,(\mathrm{NUMSTU})_j & \qquad\qquad u_{0j} = N\,(0, \tau_{00}) \\
+ \gamma 04\,(\mathrm{TCHSEX})_j + \gamma 05\,(\mathrm{TCHAGE})_j + u_0 &
\end{aligned}
$$

$\beta_{1j} = \gamma_{10}$

$\beta_{2j} = \gamma_{20}$

$\beta_{3j} = \gamma_{30}$

where

β_{1j} represents the STSEX covariate effect;

β_{2j} represents the STSES covariate effect;

β_{3j} represents the STAGE covariate effect;

γ_{00} represents the mean mathematics achievement score for a student with (1) SES equal to the class mean SES and (2) attending a class with mean SES equal to the grand mean;

γ_{01} represents the average effect of class mean SES above the grand mean SES;

γ_{02} represents the average effect of teacher experience on student achievement;

γ_{03} represents the average effect of number of student in a class on student achievement;

γ_{04} represents the average effect of teacher sex on student achievement;

γ_{05} represents the average effect of teacher age on student achievement;

γ_{10} represents the average STSEX-achievement slope;

γ_{20} represents the average effect of student SES above the class mean SES;

γ_{30} represents the average STAGE-achievement slope.

Model 3: Intercept-as-outcomes model. The same student-level model was employed here as in Model 2, but adjusted class mean mathematics achievement (β_{0j}) was further modeled as a function of OTL variables at the classroom level. Different validated OTL variables were added to the model one at a time.

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,(\mathrm{STSEX})_{ij} + \beta_{2j}\,[(SES)_{ij} - (\overline{SES})_{j}] + \beta_{3j}\,(\mathrm{STAGE})_{ij} + r_{ij} \qquad r_{ij} = N(0, \sigma^2)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}[(\overline{SES})_{j} - (\overline{SES})_{..}] + \gamma_{02}\,(\mathrm{TCHEXP})_{j} + \gamma_{03}\,(\mathrm{NUMSTU})_{j} \qquad u_{0j} = N(0, \tau_{00})$$

$$+ \gamma_{04}\,(\mathrm{TCHSEX})_j + \gamma_{05}\,(\mathrm{TCHAGE})_j + \gamma_{06}\,(\mathrm{OTL})_j + u_0$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

where

γ_{06} represents the effect of OTL on student achievement after taking all student- and classroom-level covariates into account.

Model 4: Cross-level interaction model. Model 4 includes a cross-level interaction between class-level variable, OTL, and student-level variable, $(SES)_{ij}$. Given the literature body documenting the growing inequality in academic achievement between children from low- and high-income families in Japan (Kariya, 2010; Kariya & Shimizu, 2004; Mimiduka, 2007), this model answers the question: Does OTL moderate the effect of students' SES on their achievement?

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,(\mathrm{STSEX})_{ij} + \beta_{2j}\,[(SES)_{ij} - (\overline{SES})_j] \qquad r_{ij} = N\,(0, \sigma^2)$$
$$+ \beta_{3j}\,(\mathrm{STAGE})_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}[(\overline{SES})_j - (\overline{SES})_{..}]$$
$$+ \gamma_{02}\,(\mathrm{TCHEXP})_j + \gamma_{03}\,(\mathrm{NUMSTU})_j \qquad u_{0j} = N\,(0, \tau_{00})$$
$$+ \gamma_{04}\,(\mathrm{TCHSEX})_j + \gamma_{05}\,(\mathrm{TCHAGE})_j + \gamma_{06}\,(\mathrm{OTL})_j + u_0$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}\,(\mathrm{OTL})_j$$

$$\beta_{3j} = \gamma_{30}$$

where

γ_{21} represents the interaction effect of OTL and student-level SES on achievement.

Chapter IV

RESULTS

This chapter presents the findings from the qualitative and the quantitative analytic phases. Results are presented by curriculum level in three major sections: *the intended curriculum, the implemented curriculum,* and *the obtained curriculum.* Each section is then organized as the following subsections:

1. a brief recapitulation of research question(s), analytic methods, and data sources;
2. overall results in tables accompanied with and descriptions; and
3. summary of key findings and interpretations.

Intended Curriculum **Under** ***Yutori*****: Findings of Archival Analysis of Policy Documents and Results of Descriptive Analysis of TIMSS Curriculum Questionnaires, 1999-2011**

Recapitulation of Research Question, Analysis Methods, and Data Sources

Two research questions were explored to trace the intended curriculum under *yutori.* First, I content-analyzed archival data published by MEXT to present the evolution of *yutori* reform policy trajectories from beginning (2002) to end (2011). Next, I presented the change in the Japanese 8th grade mathematics curriculum, as reported by the NRC on the TIMSS Curriculum Questionnaire, to look for disconfirming evidence in the *intended curriculum* at the national level.

Research Question One. Research Question One asked: Based on a content analysis of white papers published by MEXT between 1999 and 2011, (a) what was the *intended curriculum* as given by MEXT and (b) why did the motivations and intentions of MEXT change through the pre-*yutori*, mid-*yutori*, and post-*yutori* time periods? To answer this question, I selected documents that were available in online, open-access archival repositories of MEXT (http://www.mext.go.jp/en/publication/whitepaper/index.htm). These documents were annual white papers published by MEXT that introduced new and important policies in the areas of education, science and technology, sports, and culture for their year of publication. I content-analyzed yearly white papers published between 1999 and 2011 to map out the evolution of *yutori* policy objectives and actions during this time period, as intended by the Japanese government. My document analysis was aimed at coding and categorizing the yutori-related action of MEXT and other government agencies.

The codes I used to identify salient themes are presented in Table 9. Then I present my overarching findings in a concept map that depicts the intentions of *yutori* reform policies and how those intentions and policies evolved due to the development of other events over time.

Table 9

Example of Codebook Entry for Document Analysis

Theme	Code Name	Definition	Example of a Segment of Text From Study
Intended *yutori* reform policies	Yutori	Any evidence referring to *yutori* policies in an effort to foster independent learning in students.	“The basic concept is to cultivate a “Zest for Living” in students, such as thinking and learning for themselves, through liberal, flexible and comfortable school life under the comprehensive five-day school week” (MEXT, 1999).
Rollback of *yutori* reform policies	Rollback	Any evidence referring to changes made to undo intended *yutori* reform policies	“In the report, the philosophy “Zest for Living” is given importance, and rather than a dichotomy between “room to grow” and cramming,” there is a need to securely establish basic and fundamental knowledge and skills…Here, the proposal is to increase the class hours for Japanese language, math and arithmetic, social studies, and foreign language classes in elementary and lower secondary school to enable the firm acquisition of content easy for children to grasp through repetitive learning using knowledge and skills” (MEXT, 2007).
Adjustment to *yutori* reform aims	Aims	Any evidence referring to adjustments made to the intended *yutori* reform aims	“Content that is not specified (in the Course of Study) may be taught if a school thinks that there is a special need for it. As has been the case previously, the content specified in the Course of Study is thus the minimum standard that it has to be taught to every student” (MEXT, 2001).
Related changes instigated by international assessment results	ILSA	Any evidence referring to the development of new policies instigated by international assessment results	“According to the results of international assessments of academic achievement (PISA 2003 and TIMSS 2003) announced at the end of 2004, it indicated that the academic ability of Japanese children overall ranked high internationally, although their reading comprehension was declining.... To address these problems, the Central Council for Education issued a report in October 2005 entitled “Redesigning Compulsory Education for a New Era.” “To verify the results of education and ensure quality, this report recommended implementation of a national assessment of children’s academic achievement and degree of understanding and indicated the direction in which this national academic achievement assessment should proceed” (MEXT, 2007).

Research Question Two. Research Question Two asked: Based on the perceptions of National Research Coordinators as reported on the TIMSS surveys in 1999, 2003, and 2007, to what extent did the intended national mathematics curriculum change in terms of content coverage as *yutori* reforms were implemented between 2003-2007 in Japan? The NRCs were officials who oversaw educational policies and practices at the national or regional level. The NRCs were asked to indicate the array of TIMSS topics included in their country's intended curriculum in the 4th and 8th grades on the TIMSS Curriculum Questionnaire. To answer this research question, I reviewed the TIMSS International Mathematics Reports for 1999, 2003, and 2007 and extracted the intended 8th grade TIMSS mathematics topics, as provided by the NRCs. I then merged the three rounds of descriptive data to compare and contrast the 8th grade mathematics curriculum intended by MEXT.

Overall Results in Tables and Descriptions

Research Question One Results. Table 10 presents the themes extracted from the forewords of white papers published by MEXT from 1999 to 2011 (chronologically). Table 10 also indicates whether any references to *yutori* reforms were included in the white papers. A key finding was that except in 2001 and 2002, there was a yearly change in the leadership of MEXT, with a corresponding change in the theme of the white papers. *Yutori* reforms or their components were mentioned in all reports except in 2003 and 2004.

Table 11 presents a detailed description of *yutori* reforms and its core elements as reported in MEXT white papers. Four components of the reform were reflected in the white papers: the implementation of a 5-day school week, modification to the Course of Study, introduction of the Integrated Studies (IS) courses, and expansion of elective course offerings. Together, these components emphasized the principle that students need a liberal, flexible, and comfortable school life to develop their individuality (MEXT, 2000). However, the *yutori* reforms were revised several times from 2005 till its eventual termination in 2011 (see Table 12).

After reviewing the white papers, I found they could be clustered into three groups (1999-2002, 2003-2004, and 2005-2011), with each group reflecting similar themes. The papers for the pre-*yutori* period (1999-2002) reflected the initiation of *yutori* reforms; the papers discussed the intentions the key components of the reform. The papers for the mid-*yutori* period (2003-2004) did not reflect *yutori* reforms at all. The papers for the post-*yutori* period (2005-2011) reflected a rollback of *yutori* reforms; these papers discussed the changes made to the *yutori* reform components.

In analyzing the white papers, I found the main similarity in *yutori* reform intent over time to be the continuation of the 5-day school week policy. Because this policy was viewed as critical for students' experiential learning, even in the post-*yutori* period, it was maintained after significant revisions were made to *yutori* reforms in 2007:

> The five day school week policy is stipulated to be continued as this is a social system that was in stages over a long period of time through the cooperation of schools, homes, and communities under the basic philosophy of raining children by sharing roles in the overall society. Further, cooperation with the community is necessary to provide a variety of experiential learning activity opportunities on Saturday, within the five day school week policy, to children who desire them. (MEXT, 2007, p. 36)

I found the main discrepancies in the intentions for *yutori* policies over time to be the following:

1. **MEXT's conflicting definition of *zest for living***. In the pre-*yutori* period, *zest for living* embodied developing self-directed learning in students, yet it prioritized academic ability in the *post-yutori* period. This suggests MEXT's effort in making an appeal to reassure the importance of students' acquisition of basic academic abilities in response to public concern about declining student achievement on ILSA programs. The following quotes illustrate the contradictions.

    a. Pre-*yutori*:

    > The new Course of Study, based on the aforementioned state of children's learning, aims to realize individually targeted teaching, instead of one-way teaching of mere knowledge. With this fundamental goal in mind, it carefully selects educational content so that every child can acquire fundamentals and basics, and tries to develop a "zest for living," such as an ability to learn and think on his or her own. (MEXT, 2001, p. 3)

    b. Post-*yutori*:

    > Following the revision of the Fundamental Law of Education, MEXT revised the Course of Study to establish the measures for realizing the philosophy of "Zest for Living" such as academic ability, generous spirit health and physical strength of children. (MEXT, 2008, p. 1)

2. **The Japanese Government's conflicting position on academic pressure and competition**. MEXT originally proposed the *yutori* reform policy under the premise of relieving students' academic stress, yet the same agency turned around and embraced academic competition 5 years later. Perhaps this was due to the frequent change in the leadership of MEXT, as shown in Table 9.

Below are two excerpts from the Minister of Education's Foreword for a pre-*yutori* and a post-*yutori* white papers.

a. Pre-*yutori*:

> Increased competition in examinations has resulted in school education being reduced to a form in which academic knowledge is one-sidedly instilled in students, thus leading to the neglect of education and activities that cultivate thinking faculties, creativity, and an enriched humanity. (MEXT, 1999, p. 1)

b. Post-*yutori*:

> I hope that through educational reforms for restoring people's vitality, the adverse effects of the principle of free competition will be reduced and that people of moral character will build Japan as a nation with dignity. (MEXT, 2006, p. 1)

When I organized the coded *yutori*-related events over time in a concept map, the interconnected events strongly suggested that MEXT's conflicting stances may have been instigated by the ILSA reports. As illustrated in Figure 5, MEXT announced the *yutori* reform policies in 1998 on the recommendation of the Central Council for Education. Between 1998 and the eventual implementation of *yutori* in 2002, there were hardly any changes to MEXT's reform plan. Except in 2001, when the PISA 2000 results were released, *yutori* critics had questioned MEXT about whether Japanese students would continue to perform at the top level once reforms began, as they had on the PISA assessments in 2000. In response, MEXT issued an official statement to emphasize that "the content specified in the Course of Study is the minimum standard" (MEXT, 2001, Chapter 3, Section 2.2.4) and "the content not specified (in the Course of Study) may be taught if a school thinks that there is a special need for it" (MEXT, 2001, Chapter 3, Section 2.2.4). The cause-and-effect link of this particular set of events is displayed as a red dashed line in Figure 5. Then MEXT proceeded to implement *yutori* reforms as intended, which consisted of four main components (highlighted in a black box on Figure 5) in all public schools from kindergarten to 9th grade.

With the release of the TIMSS 2003 and PISA 2003 results in 2004, MEXT was under pressure again to address the declining performance of Japanese students. As a result, MEXT (2005) established the National Assessment of Academic Ability to "guarantee education quality by setting clear goals and examining outcomes" (Chapter 1, p. 3). This event is displayed as green dashed line in Figure 5. The TIMSS 2003 and PISA 2003 results also triggered the then-prime minister of Japan, Shinzo Abe, to create the Education Rebuilding Council (ERC) in 2006. In its second report released in June 2007, the ERC recommended major revisions to the Course of Study to improve academic ability. MEXT immediately complied with that recommendation and announced a revised Course of Study in 2007. The new Course of Study for mathematics and science was implemented in all elementary and junior high schools in 2008; the Course of Study was implemented at full scale in 2010. This chain of events is presented as a green solid line in Figure 5.

The release of the PISA 2006 and TIMSS 2007 results were not explicitly mentioned in the white papers and they did not seem to trigger any reactions to *yutori* reform policies. The *yutori* era officially ended in 2011 when prime minister Abe announced a new set of reforms to supersede *yutori.*

Table 10

Themes of White Papers Published by MEXT Between 1999-2011

Year	Theme	Report Mentioning *yutori* Reforms	Minister of Education Who Published the White Papers
1999	Ensuring all children grow up with good health in a life-enriching environment	Yes	Hirofumi Nakasone
2000	Actualizing a culturally-oriented nation	Yes	Tadamori Oshima
2001	Implementing the Educational Reform Plan for the 21st Century	Yes	Astuko Toyama
2002	Promoting educational reform that cultivates independence and creativity	Yes	Astuko Toyama
2003	New developments in higher education reform	No	Takeo Kawamura
2004	Development of healthy minds and bodies	No	Nariaki Nakayama
2005	Providing an enriched education for all children	Yes	Kenji Kosaka
2006	Rebuilding education and promoting culture and the arts	Yes	Bunmei Ibuki
2007	Revising educational reforms to usher in a new era of education	Yes	Kisaburo Tokai
2008	Comprehensive promotion of education policy	Yes	Ryu Shionoya
2009	Fostering people and knowledge through the promotion of education and culture, sports, science, and technology	Yes	Katsuo Kawabata
2010	Creating a Sport Nation and fostering future directions of education policy measures	Yes	Yoshiaki Takaki
2011	Restoration following the earthquake of 2011	Yes	Masaharu Nakagawa

Table 11

Key Components of Yutori Reforms Derived From MEXT White Papers

Key Components of the *yutori* Reforms	Description	Goals/Aims	White Papers (Year)
A. Shortened school week	Implementation of 5-day school week	To cultivate zest for living in students by reducing excessive studying	1999 2000 2001 2002
B. Modification to the Course of Study [a]	1. Up to 30% reduction of curricular content in core academic subjects at elementary and junior high schools via: -Moving advanced contents to higher grades -Reducing content overlap between subjects and grades 2. Reduction of annual class hours for core subjects	To rekindle students' interest in academics To give students ample time to acquire basic skills	1999 2000 2001 2002
C. Introduction of the Integrated Studies (IS) courses	1. Establishment of interdisciplinary classes which could be designed and offered at each school's own discretion 2. On average, there were 2-4 hours of IS classes per week for 3rd graders and above (out of 28-hour school week) 3. No tests and no grades were given for IS classes	To nurture self-initiated learning To foster an independent and creative attitude towards problem-solving	1999 2000 2001 2002
D. Expansion of the range of elective courses	1. Initially introduced in the 1989 curriculum 2. Weekly hours for electives rose to between 1.5 to 5 hours in junior high schools 3. Schools could offer electives that expanded on concepts introduced in the existing curriculum	To encourage autonomous learning and individuality To offer students more opportunities for supplemental learning in core subjects	1999 2000 2001 2002

[a] Course of Study outlines a series of courses that all students are required to complete before they can move on to the next grade level.

Table 12

Revisions to Key Components of Yutori Reforms Outlined in MEXT White Papers

Key Components of the *yutori* Reforms	Description of the Change	Revised Goals/Aims	White Papers (Year)
A. Five-day school week	No change	N/A	N/A
B. Course of Study	1. 11% to 29% increase in Japanese, math, and science in class content and hours for elementary and junior high schools 2. One hour increase in weekly class hours for junior high schools 3. Establishment of Nationwide Assessment of Academic Ability which would be administered annually to all students in the six years of elementary school and the third year of junior high school.	To bolster academic ability To ensure education quality by setting clear goals and examining outcomes	2005 2007 2008 2009 2010 2011
C. IS courses	Reducing IS course to no more than two hours per week for elementary schools and junior high schools	To make time for the new Course of Study within the existing 5-day school week	2007 2008 2009 2010 2011
D. Elective courses	Reducing elective courses to as much as zero hours per week at junior high schools	To uncomplicated the burden on schools for offering electives and IS classes at the same time	2007 2008 2009 2010 2011

Note: N/A indicates no changes were made.

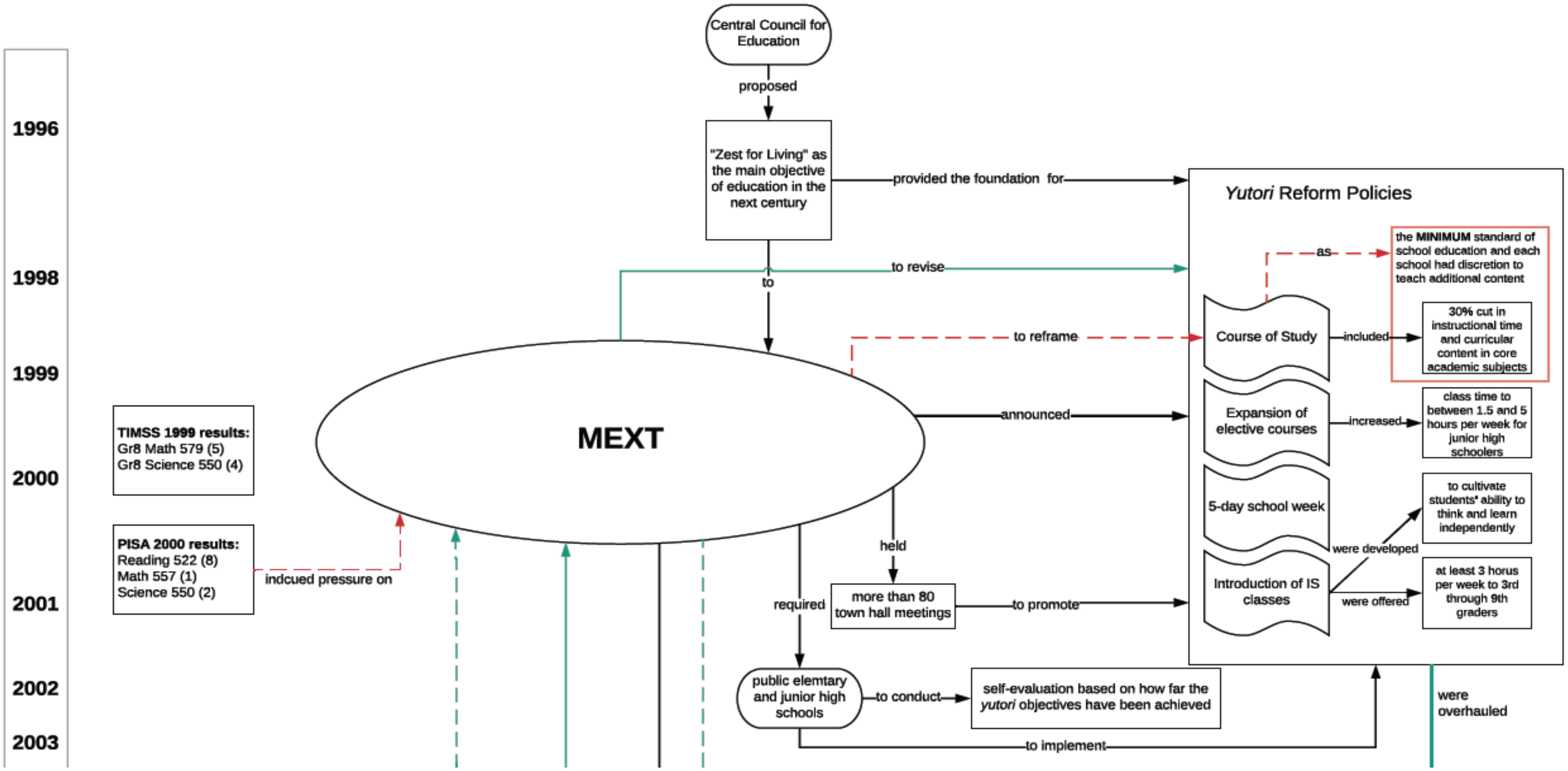
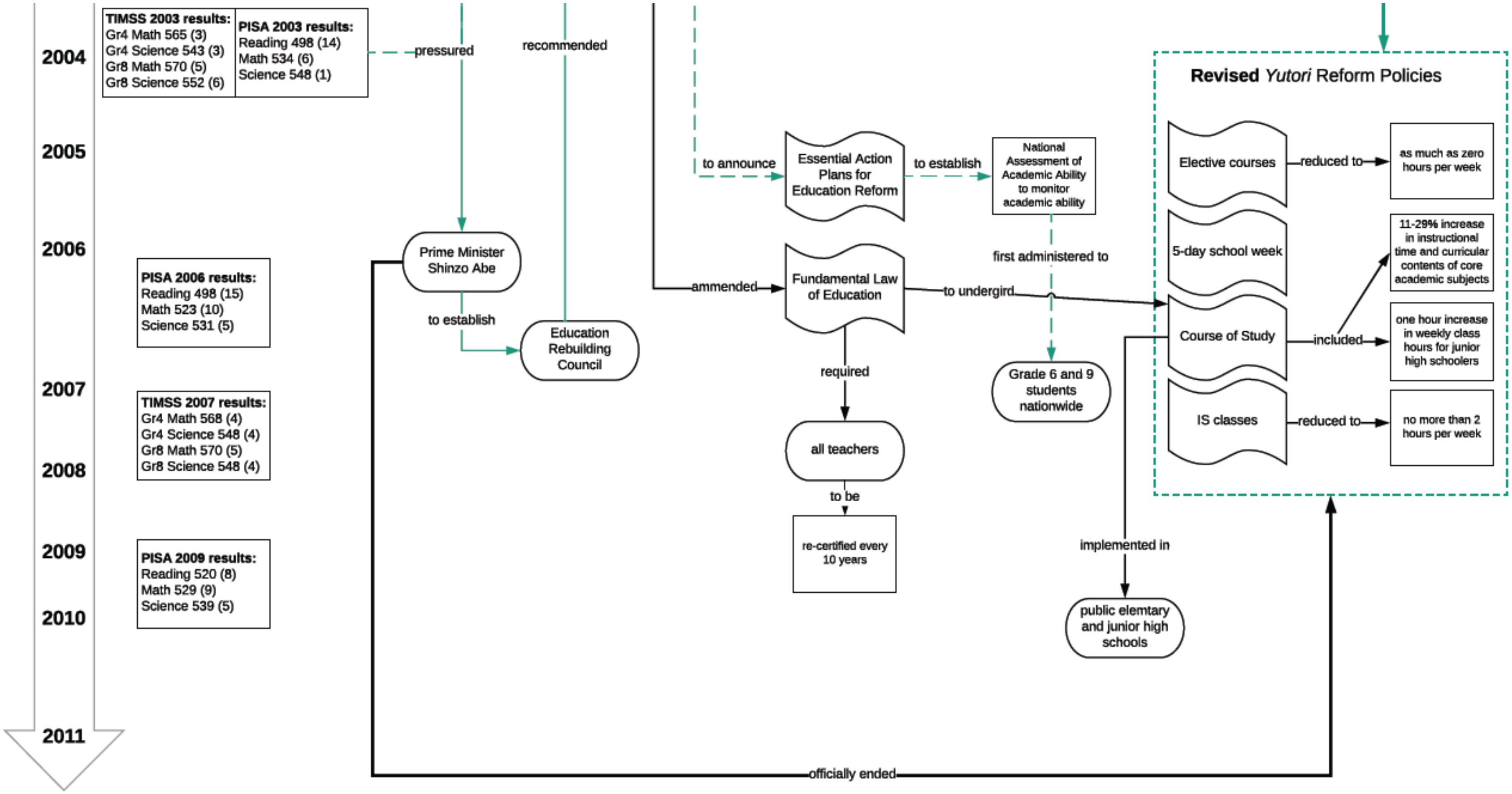

1996
1998
1999
2000
2001
2002
2003
Central Council for Education
proposed
"Zest for Living" as the main objective of education in the next century
provided the foundation for
to revise
to
to reframe
MEXT
announced
TIMSS 1999 results:
Gr8 Math 579 (5)
Gr8 Science 550 (4)
PISA 2000 results:
Reading 522 (8)
Math 557 (1)
Science 550 (2)
indcued pressure on
held
more than 80 town hall meetings
to promote
required
public elemtary and junior high schools
to conduct
self-evaluation based on how far the *yutori* objectives have been achieved
to implement
Yutori Reform Policies
as
the MINIMUM standard of school education and each school had discretion to teach additional content
Course of Study
included
30% cut in instructional time and curricular content in core academic subjects
Expansion of elective courses
increased
class time to between 1.5 and 5 hours per week for junior high schoolers
5-day school week
were developed
to cultivate students' ability to think and learn independently
Introduction of IS classes
were offered
at least 3 horus per week to 3rd through 9th graders
were overhauled

Figure 5. MEXT's *yutori* reforms implementation timeline by year

(continued on next page)



Results for International Large Scale Assessment (ILSA) are presented in red font.

Research Question Two Results. From my review of the TIMSS Curriculum Questionnaire from 1999, 2003, and 2007, I compiled a comparison of the intended 8th grade mathematics curriculum over time (see Table 13). Focusing on the common TIMSS mathematics topics across the three rounds, I found that there was almost no difference in the TIMSS mathematics topics covered in 1999, 2003, and 2007, as reported by the Japanese NRC. I found that the arithmetic mean, median, and mode suggested that topics were offered pre-*yutori* (indicated by a solid circle in 1999), but were not covered during post-*yutori* periods (indicated by an empty circle in 2003 and 2007). These results suggested that the intended 8th grade mathematics curriculum remained virtually unchanged even after the implementation of *yutori* reforms. However, this analysis was limited by the number of common mathematics topics covered in these three rounds of the TIMSS.

Summary of Key Findings

Since its introduction in 1998, *yutori* reforms underwent several rounds of changes. Almost all changes were triggered by the ILSA reports (i.e., the PISA and TIMSS reports), either directly or indirectly. When the ILSA results suggested that the academic attainment of Japanese students was slipping, MEXT would take direct actions to respond. Additionally, MEXT was also under pressure from the prime minister’s office to address the declining ILSA results. With pressure mounting, MEXT’s actions over time went from re-emphasizing the objectives of *yutori* to overhauling major components of the reform. Perhaps the frequencies of change to *yutori* were further exacerbated by the annual change in the minister of education position. With a high turnover at the top, it

Table 13

TIMSS Topics Covered in the Intended and the Implemented Curricula for 8th Grade Mathematics

Major Topics	Subtopics	TIMSS 1999		TIMSS 2003		TIMSS 2007	
		Intended Curriculum	Implemented Curriculum	Intended Curriculum	Implemented Curriculum	Intended Curriculum	Implemented Curriculum
Fractions and Number Sense	Whole numbers - including place values, factorization and operations (+,-,x,/)	●	99%	●	100%	●	96%
	Understanding and representing common fraction	●	98%	●	98%	●	98%
	Computations with common fractions	●	100%	●	100%	●	99%
	Understanding and representing decimal fraction	●	98%	●	98%	●	98%
	Computations with decimal fractions	●	100%	●	100%	●	98%
	Relationships between common and decimal fractions ordering of fractions	●	99%	-	-	●	97%
	Rounding whole numbers and decimal fractions	●	92%				
	Estimating the results of computations	●	89%	●	97%	●	96%
	Number line	●	100%	-	-	-	-
	Whole number powers of integers✢	●		-	-	-	-
	Computations with percentages and problem involving percentages	●	100%	●	98%	●	98%
	Simple computations with negative numbers	●	100%	-	-	-	-
	Square roots (of perfect squares less than 144), small integer exponents	○	14%	-	-	-	-
	Prime factors, highest common factor, lowest common multiple, rules for divisibility✢	○		-	-	-	-
	Sets, subsets, union, intersection, venn diagrams✢	○		-	-	-	-
	Rate problems✢	●		-	-	-	-
	Concepts of ratio and proportion; ratio and proportion problems	●	97%	●	91%	●	87%
	Integers including words, numbers, or models	-	-	●	99%	●	100%

✢ Topics not included in Teacher Questionnaires. The implemented curriculum column reflected the percentages of students who were taught a particular topic, as reported by their teachers. A dash (-) indicates comparable data are not available. The intended curriculum reflected whether a topic was included in the national curriculum using the following symbols.

● denotes topics were included in the intended curriculum to be taught to all or almost all students (at least 90%)

○ denotes topics were not included in the intended curriculum

Table 13 (continued)

		TIMSS 1999		TIMSS 2007		TIMSS 1999	
Major Topics	Subtopics	Intended Curriculum	Implemented Curriculum	Intended Curriculum	Implemented Curriculum	Intended Curriculum	Implemented Curriculum
Measurement	Units of measurement, standard metric units	●	90%	●	96%	-	-
	Reading measurement instruments	●	84%	●	90%	-	-
	Estimates of measurement, accuracy of measurement	●	66%	○	43%	-	-
	Conversions of units between measurement systems✢	●		●	91%	-	-
	Perimeter and area of simple shapes - triangles, rectangles, and circles	●	99%	●	95%	●	96%
	Perimeter and area of combined shapes	●	78%	●	70%	-	-
	Volume of rectangular solids i.e., Volume = length x width x height	●	98%	●	95%	●	96%
	Volume of other solids (e.g., pyramids, cylinders, cones, spheres)✢	●		-	-	-	-
	Computing with measurements (+,-,x,/)✢	●		●	61%	-	-
	Scales applied to maps and models	●	84%	-	-	-	-
	Estimations of length, circumference, area, volume, weight, time, angle, and a speed in problem situations	-	-	●	89%	-	-
Data Representation, Analysis, and Probability	Collecting and graphing data from a survey✢	●		○	19%	-	-
	Representation and interpretation of data in graphs, charts, and tables	●	43%	●	55%	●	52%
	Arithmetic mean	●	38%	○	9%	○	13%
	Median and mode✢	●		○	9%	○	13%
	Simple probabilities - understanding and calculations	○	3%	●	33%	●	51%
	Sources of error in collecting and organizing data	-	-	○	12%	●	12%
	Evaluating interpretations of data with respect to correctness and completeness of interpretation	-	-	○	4%	-	-
	Interpreting data sets	-	-	○	6%	○	17%
	Organizing a set of data by one or more characteristics using a tally chart, table or graph	-	-	●	24%	●	48%

✢ Topics not included in Teacher Questionnaires. The implemented curriculum column reflected the percentages of students who were taught a particular topic, as reported by their teachers. A dash (-) indicates comparable data are not available. The intended curriculum reflected whether a topic was included in the national curriculum using the following symbols:

● denotes topics were included in the intended curriculum to be taught to all or almost all students (at least 90%)

○ denotes topics were not included in the intended curriculum

Table 13 (continued)

Major Topics	Subtopics	TIMSS 1999 Intended Curriculum	TIMSS 1999 Implemented Curriculum	TIMSS 2003 Intended Curriculum	TIMSS 2003 Implemented Curriculum	TIMSS 2007 Intended Curriculum	TIMSS 2007 Implemented Curriculum
Geometry	Using the chances of a particular outcome to solve problems	-	-	-	-	●	58%
	Coordinates of points on a given straight line	●	99%	-	-	-	-
	Simple two-dimensional geometry - angles on a straight line, parallel lines, triangles and quadrilaterals	●	97%	●	98%	●	100%
	Congruence and similarity	●	98%	●	97%	●	99%
	Angles - (acute, right, supplementary, etc.)$^{+}$	●		●	91%	●	98%
	Pythagorean theorem (without proof)$^{+}$	○		○	2%	○	4%
	Symmetry and transformations (reflection and rotation)	●	98%	●	88%	●	99%
	Visualization of three-dimensional shapes	●	82%				
	Geometric constructions with straight-edge and compass^{+}	●		●	92%	●	93%
	Regular polygons and their properties - names (e.g., hexagon and octagon), sum of angles, etc.$^{+}$	●		●	94%	●	100%
	Proofs (formal deductive demonstrations of geometric relationships)$^{+}$	●		-	-	-	-
	Sine, cosine, and tangent in right-angle triangle^{+}	○		-	-	-	-
	Nets of solids^{+}	●		-	-	-	-
	Translation, reflection, rotation, and enlargement	-	-	○	67%	○	79%
	Relationship between 2-dimensional and 3-dimensional shapes	-	-	●	51%	●	89%
	Similar triangles and recall their properties	-	-	○	4%	○	7%
	Properties of angle bisectors and perpendicular bisectors of lines	-	-	●	98%		
	Measures of irregular or compound areas	-	-	-	-	●	56%
	Measurement, drawing, and estimation of the size of angles, the lengths of lines, areas, and volumes	-	-	-	-	●	95%

+ Topics not included in Teacher Questionnaires. The implemented curriculum column reflected the percentages of students who were taught a particular topic, as reported by their teachers. A dash (-) indicates comparable data are not available. The intended curriculum reflected whether a topic was included in the national curriculum using the following symbols:

● denotes topics were included in the intended curriculum to be taught to all or almost all students (at least 90%)

○ denotes topics were not included in the intended curriculum

Table 13 (continued)

Major Topics	Subtopics	TIMSS 1999		TIMSS 2003		TIMSS 2007	
		Intended Curriculum	Implemented Curriculum	Intended Curriculum	Implemented Curriculum	Intended Curriculum	Implemented Curriculum
Algebra	Number patterns and simple relations	●	94%	-	-	-	-
	Writing expressions for general terms in number pattern sequence✛	●		●	77%	●	71%
	Translating from verbal descriptions to symbolic expressions✛	●		-	-	-	-
	Simple algebraic expressions	●	100%	●	95%	●	98%
	Evaluating simple algebraic expressions by substitution of given value of variables✛	●				●	99%
	Representing situations algebraically; formulas	●	98%	●	93%	●	94%
	Solving simple equations	●	100%	-	-	●	94%
	Solving simple inequalities	●	99%	-	-	●	94%
	Solving simultaneous equations in two variables✛	●		-	-	●	94%
	Interpreting linear relations✛	●		-		-	-
	Using the graph of a relationship to interpolate/extrapolate✛	●		●	97%	-	-
	Proportional, linear, and nonlinear relationships	-	-	●	92%	-	-
	Sums, products, and powers of expressions containing variables	-	-	●	96%	●	92%
	Equivalent representations of functions as ordered pairs, tables, graphs, words or equations	-	-	-	-	●	91%
	Evaluating expressions for given numeric value	-	-	-	-	●	100%

✛ Topics not included in Teacher Questionnaires. The implemented curriculum column reflected the percentages of students who were taught a particular topic, as reported by their teachers. A dash (-) indicates comparable data are not available. The intended curriculum reflected whether a topic was included in the national curriculum using the following symbols:

● denotes topics were included in the intended curriculum to be taught to all or almost all students (at least 90%)

○ denotes topics were not included in the intended curriculum

may have been extremely difficult for the leadership to maintain control over implementation of such significant reforms. Even though *yutori* reforms constituted up to a 30% cut in mathematics curriculum in elementary and junior high schools, I found only a few differences (specifically in six areas, see Table 13) in terms of topic coverage in the pre- and post-*yutori* years for the intended mathematics curriculum for 8th graders, as provided by the TIMSS Curriculum Questionnaire.

***Implemented Curriculum* Under *Yutori*: Findings of Analysis of Interview Data, Validations of TIMSS Construct Measures of OTL, and Results on OTL in Mathematics Based on TIMSS Teacher Questionnaires**

Recapitulation of Research Questions, Analysis Methods, and Data Sources

Three research questions (Research Question 3-5) were explored to describe the implemented curriculum under *yutori* reforms. First, I content-analyzed interview data collected from two semi-structured interviews researchers who have studied *yutori* reforms extensively. Next, I empirically validated OTL construct measures following Stevens' (1996) OTL framework. Last, I used the validated OTL measures and conducted quantitative analysis to compare mean differences in OTL across the three rounds of the TIMSS.

Research Question Three. Research Question Three asked: From two school-based researchers' observations of Japanese junior high schools during the *yutori* reform period: (a) what roles did teachers and schools play in implementing *yutori* reform directives, (b) what support and barriers did schools and teachers face as they attempted to follow *yutori* reform guidelines, and (c) what were other contextual factors affecting the implementation of *yutori* reforms in Japanese schools and classrooms? To answer this question, I interviewed two researchers who had studied *yutori* reforms extensively. On February 28, 2018, I interviewed the first researcher (hereinafter referred to as "Researcher A") who had conducted fieldwork over a 6-year period (2003-2009) which included: (a) interviews with 40 teachers, administrators, and MEXT officials regarding their opinions and experiences related to *yutori* reforms; and (b) a full year of in-depth ethnographic study in three elementary schools and three junior high schools located in Northwest Japan. On March 1, 2018, I interviewed the second researcher (hereinafter referred to as "Researcher B") who had conducted a longitudinal, multisite ethnographic study of two junior high school located in West-central Japan on many visits between 1994 and 2007, and then supplemented the ethnographic study with survey research and analysis of documents issued by the Japanese government and major media. To further support their recount, both researchers referred me to their publications on *yutori* reforms.

Research Question Four. Research Question Four asked: To what extent is Stevens' multidimensional framework suggesting four interrelated OTL constructs, upheld in the TIMSS 1999, 2003, and 2007 survey data from samples of participating Japanese 8th mathematics teachers? The purpose of this research question was to derive and validate OTL measures using theory in order to examine changes in the implementation levels of OTL at the pre-, mid-, and post-*yutori* reform periods in the following quantitative phase. As indicated in Chapter III, the Process Model (Chatterji, 2003, in press) is an iterative instrument design and validation methodology that relies on different but relevant kinds of validity evidence to assure the psychometric quality and meaningfulness of construct measures to best meet the inferential needs of users. Chapter III also provided the domain specifications and theoretically-derived indicators for OTL subdomains based on Stevens' (1996) framework, alongside the TIMSS teacher survey items that I matched based on my own judgment (see Table 7 in Chapter III).

To address this particular research question, I used the Process Model to validate the OTL measures derived from TIMSS teacher surveys. I gathered and evaluated evidence of content validity, internal scale structure based on principal components analysis (PCA), convergent validity, and scale score reliability. To content-validate the items independently, I asked two external experts to evaluate OTL items for a reasonable match for the specified indicators of the multidimensional OTL framework (Stevens, 1996) in terms of representativeness and relevance. Then I conducted PCA and estimated convergent validity and scale score reliability to confirm and evaluate the quality of the derived OTL construct measures.

Research Question Five. Research Question Five asked: Using the validated OTL measures per Stevens' framework, to what extent did the *yutori* curricular reforms affect changes in OTL levels over time in 8th grade mathematics classes, as given by the TIMSS teacher survey data from 1999, 2003, and 2007?

To address this question, I first reviewed the TIMSS Teacher Questionnaires to extract information corresponding to the theoretically-derived *content coverage* and *content emphasis* aspects of Stevens' OTL framework. As indicated previously, the *content coverage* and *content emphasis* items were descriptively based and contained categories that were measured on a nominal scale. As such, these two scales could not be compared across time using inferential tests. Thus, I presented the changes in *content coverage* and *content emphasis* over time using descriptive analysis of the TIMSS teacher surveys.

Next, I merged the three data sets and then conducted a one-way ANOVA followed by post-hoc Tukey HSD tests to see if there were significant mean differences in the two validated OTL measures—*quality of instructional delivery* OTL and *instructional time*—across three consecutive rounds of the TIMSS. All assumptions for one-way ANOVA were checked prior to conducting the analysis. Recall that *instructional time* is the single-item variable that I provisionally kept with literature support.

Overall Results in Tables and Descriptions

Research Question Three results. The five main themes that emerged from the researcher interviews were: (a) lack of implementation fidelity to the intent of *yutori* reforms, (b) individual barriers to *yutori* reforms, (c) institutional barriers to *yutori* reforms, (d) pilot schools as a facilitator to *yutori* reforms, and (e) growing achievement gaps between low and high SES students as an unintended consequence of *yutori* reforms. Findings on each theme are elaborated with representative quotes below.

Lack of implementation fidelity of yutori reforms. Teachers' responsibilities progressively increased after the official start of *yutori* in 2002. Teachers reported difficulties planning and teaching the new IS courses dedicated to student-centered, experiential learning in elective subjects (MEXT, 2002). Because MEXT chose not to publish textbooks for IS or elective courses, teachers had to develop their own curriculum and materials. Researcher B cited a junior high school teacher who captured the difficulties teachers faced in developing their own materials:

> Elective subjects—well, they are villainous. They need amazing amounts of labor. The amount of time you need just to prepare for one lesson is incredible, over and above what is needed for ordinary subjects. It's a huge burden. Frankly, if they vanish, it'll be a relief.

According to Researcher A, teachers continued to feel pressure to cover the contents of the pre-*yutori* curriculum in core academic subjects:

> Teachers were conscious of the topics that had been covered on entrance exams in the past, so if they noticed that the new curriculum didn't attach enough importance to a particular concept, a lot of times they would use old textbooks or supplementary activities to cover that material.

Although the Course of Study provided the number of hours that should be allocated to each subject at each grade level, teachers found ways to subvert this order and use time presumably set aside for IS and elective courses for core academic subjects. Researcher A shared a quote from a junior high school teacher that vividly summarized the implementation of *yutori* reforms:

> To be honest, we don't always teach what we're supposed to. We don't have enough time in the schedule to teach math, so we use the morning IS time for math. If we don't use that time for math, we have to find other time to teach math. We can also use cleaning time for instruction—fifteen minutes a day adds another hour to the schedule.

Lastly, a core tenet of *yutori* reforms was the idea that instruction should be dynamic and student-centered. However, Researcher B reported otherwise:

> Mathematics classes usually comprised a series of short cycles: teacher explanation would be followed by the students' working on some practice exercises individually, and then the checking of the answers by the teacher.

Individual barriers to yutori reforms. Three individual teacher-level barriers to *yutori* reforms were identified: (a) lack of capacity, (b) lack of time, and (c) too much autonomy.

Lack of capacity. Lack of capacity to teach the new IS courses was a major barrier to implementation. Because of their interdisciplinary nature, IS courses demanded much more knowledge and skills than most teachers possessed. As Researcher B reported:

> Integrated studies were much more demanding of teachers than their own subject. Building up their capacity to teach it demanded considerable time, both for intensive initial training and for continuing training to maintain creativity.

Researcher A added that:

> [IS course] takes a lot of practice and experience to make it work, but these teachers had to start using it right away and that was very, very difficult for them. And so in general, the teachers needed more support at the level of staff instruction and the level of curricular materials.

Lack of time. Time constraint was another major barrier to implementation identified by both researchers. Junior high school teachers had many non-teaching duties, according to Researcher B:

> Besides teaching classes, usually for three or four hours a day, teachers were expected to engage in behavioral guidance; support and direct students in nonacademic activities, such as preparation for the sports day, cultural festival, or choral contest; supervise extracurricular clubs; make and mark periodic tests; determine and record grades; take part in various grade, school, and committee meetings; and (if a class teacher) eat lunch with their class and look at students' schedules and diaries every day.

Researcher A shared that Japanese junior high school teachers had long work hours every day:

> Junior high school teachers, middle school teachers, in Japan work very, very hard. At the schools where I was doing my research teachers almost never left campus before 7:00, often there until 10:00, 11:00 at night, and so they just simply didn't have enough time to invest in the activities that would ensure successful lessons.

Too much autonomy. Unfortunately, MEXT's strategies of entrusting schools and teachers to roll out the reform guidelines and evaluate their own reform progress became a barrier to implementation. Entrusted to self-evaluate, schools were required to submit weekly schedules and official reports to MEXT documenting their *yutori* reform progress. To avoid criticism, these reports presented an "idealized" version of the actual implementation. Researcher A provided the following evidence:

> The reports that schools submitted to MEXT did not always accurately describe what they were actually doing, so a lot of times...I mean, all the schools I visited, they were very adept at writing reports and submitting formal schedules that showed that they were following the *yutori* guidelines exactly the way MEXT wanted them, but that a lot of times wasn't what was actually going on. They could say they have a class for independent research, or whatever, or a new elective, but in actuality they would use that time for math review or something, and that wouldn't appear on the report.

The high level of trust and autonomy given to teachers also gave them the freedom to use IS and elective courses for academic subjects, as evident in the quotes provided previously to illustrate lack of implementation fidelity.

Institutional barriers to yutori reforms. Two individual barriers to *yutori* reforms were identified: lack of leadership and firm institutional priorities.

Lack of leadership. Leadership from the principals was noticeably absent in schools during the *yutori* reform period. Researcher B offered the following explanation:

> Principals are generally in post for only two or three years before being transferred to another school. It is therefore extremely difficult for them to lead change, since doing so takes time.

Researcher A concurred:

> The principals tend to be more focused on communicating to public and they take a hands-off role, so I talked to all the principals, but they weren't really actively involved.

Firm institutional priorities. Both researchers emphasized institutional priorities as a major barrier to the reform effort at each school because of the clash between the demands made by the *yutori* reforms and institutional beliefs and practices at school level. Institutional priorities reflected long-standing ideals and practices that were integrated and institutionalized at each school. These institutional priorities such as academic subject teaching, maintaining order, and caring for one's class accentuated teachers' resistance to comply to *yutori* guidelines. For example, according to Researcher A, junior high school teachers commonly believed that the primary role of junior high school is to prepare students for the next level of education:

> Instruction as well as guidance becomes firmly focused on preparation for high school examinations…. Instructors place greater emphasis on the recall of factual information—especially materials that might appear on entrance examination.

Researcher B cited the clash between non-teaching tasks and lack of time to plan for IS classes as an illustration of institutional priority:

> Clearly, club supervision was a higher priority for them than planning and preparation for integrated studies. The explanation for this was that club supervision was personally rewarding for many teachers, and it was also a strongly institutionalized activity that was perceived to be vital for students' personal development and their integration into the school.

Pilot schools as a facilitator to* yutori *reforms. Professional development via pilot schools was cited as the only support that teachers and schools received from the central governance. According to Researcher A:

> Teams of teachers did go and visit the pioneer schools and they did attend professional development workshops.... I know that there were workshops held, and even when I went in 2003 when I went to national *yutori* conferences and teachers were drawn from all over the country to learn more about it, and to learn about how they might alter their curricula to fit the goals of the *yutori*. There was professional development offered on many different levels, some nationally, some in the prefectures, and some locally.

However, this kind of professional development was not available to every teacher. Instead, schools would designate teachers from their *yutori* implementation team to attend training sessions, as Researcher B recalled:

> Somebody would be put in charge of integrated studies, for example in each school, and that person would go to all the training sessions and so on, and then it would be that persons' responsibility to organize training sessions and so on... training sessions is not exactly the right word, but you know, sessions in which the staff would discuss and learn about what to do in integrated studies…. But what tended to happen I think was that the small number of individuals who were taking part in the training sessions, they would end up as being among the few enthusiasts for the reform in school, while everybody else was essentially apathetic or hostile.

Growing achievement gap as an unintended consequence of yutori reforms. Ever since its inception, *yutori* reforms have received very negative media coverage. Researcher A described:

> Parents were very concerned about the impact of this change on their children's opportunities to get into desirable high schools, and those concerns tended to be heightened by accounts of media. There were at the time a lot of reports on TV shows and in newspapers, by pundits who were very critical of the reforms, and parents picked up on that. Although most parents didn't have a lot of concrete information, they were very worried.

Juku, which are private, afterschool tutoring institutions, capitalized on the fear of teachers and mostly parents that the *yutori* curriculum was not going to serve their students well. Researcher A offered some insights into *juku's* strategies of exploiting parents' fear of *yutori* reforms:

> While I was there I visited *juku* and interviewed *juku* operators and they shared with me publicity materials they developed, brochures, pamphlets, ads [sic] they placed in newspapers, and they frequently emphasized the dangers of *yutori* and how sending kids to *juku* could help fill those gaps and ensure that kids weren't left behind because of problems that emerged in the schools. The *juku* operators consciously publicized their ability to fill in those gaps and cover a lot of the material that had been cut to alleviate parents' concerns. And *juku* attendance did increase, although I didn't document, I can't tell you exactly how much.

Researcher B further buttressed the above inference with these words:

> These parents felt strongly that their children should make their own decisions about whether to go to *juku*. However, it was clear that the pressures that stemmed from worry about losing out to others in the competition to secure good educational credentials could be hard for both parents and children to resist.

Unfortunately, *jukus* are expensive to attend. According to Researcher B, in 1997, tuition started at ¥10,000 a month (US$83 at the 1997 exchange rate of US$1 = ¥120) for 3 hours of instruction per week. Inferring from their high costs, *jukus* were not easily accessible to students from low-SES families.

Research Question Four results. To address this question, content validation and empirical validation results of the OTL scales are presented below.

Content validation results. Experts in classroom pedagogy in mathematics, international large-scale assessment, and measurement were consulted to evaluate the items I selected from the TIMSS teacher surveys to match appropriate theoretically-derived indicators based on the literature review and Stevens' framework. Following the content-based reviews, I estimated the content validity of OTL scales using the content validity index (CVI), as proposed by Polit et al. (2007).

Table 14 presents results from the content-based reviews by the experts. Using guidelines suggested by Polit and colleagues (2007), items with kappa of 1.0 were considered having an excellent match to the proposed indicators, where items with kappa of 0 were considered having a poor fit to the proposed indicators. Five items from *content emphasis* and two items from *quality of instructional delivery* were dropped due to poor fit. The resulting OTL construct measures, which were composed of 54 items from the TIMSS Teacher Questionnaire, are shown in Table 15.

Table 14

Content Validation of TIMSS Teacher Survey Items That Match Stevens' OTL Framework

Construct	Indicators	Questions	Scale	Number of Experts	Number Giving Rating of 3 or 4	I-CVI[a]	k*[b]	Evaluation[c]
1.0 Content coverage	1.1 Teacher arranges for all students to have access to the core curriculum 1.2 Teacher arranges for all students to have access to critical subject matter 1.3 Teacher ensures that there is curriculum content and test content overlap	The following list includes the main topics address by the TIMSS math test. Check the response that describe when students in your math class have been taught each topic. ***Major topics: Number, Algebra, Geometry, Measurement, Data and Chance***	1 = Taught before this year 2 = Mostly taught this year 3 = Not yet taught	2	2	1.00	1.00	Excellent
2.0 Content exposure	2.1 The amount of time teachers allocated to covering the content 2.3 Time devoted to a subject area (i.e. math)	How many minutes per week do you teach math to your math class?	Continuous	2	2	1.00	1.00	Excellent
	2.2 Time allotted to students to learn 2.3 Time devoted to a subject area (i.e. math)	If you assign math homework, how many minutes of math homework do you usually assign your students?	1 = Less than 15 minutes 2 = 15-30 minutes 3 = 31-60 minutes 4 = 61-90 minutes 5 = More than 90 minutes	2	2	1.00	1.00	Excellent
	2.1 The amount of time teachers allocated to covering the content 2.2 Time allotted to students to learn	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on each of the following activities? •homework review •lecture-style presentation by teacher •teacher-guided student practice •re-teaching and clarification of content/procedures •student independent practice •tests and quizzes •other	Percentages	2	2	1.00	1.00	Excellent

[a] I-CVI, item-level content validity index.
[b] k* = kappa designating agreement on relevance.
[c] Evaluation criteria for kappa, using guidelines described in Polit, Beck, & Owen (2007): Fair = k of .40 to .59; Good = k of .60–.74; and Excellent = k > .74.

Table 14 (continued)

Construct	Indicators	Questions	Scale	Number of Experts	Number Giving Rating of 3 or 4	I-CVI[a]	k*[b]	Evaluation[c]
3.0 Content emphasis	3.1 Teacher selects topics within the curriculum to teacher (i.e. as a major topic, a minor review or not taught at all)	What subject matter do you emphasize the MOST in your math class? **Subjects: Number, Algebra, Geometry, Combined Algebra and Geometry, Combined Algebra, Geometry, Number, etc., Other.**	Percentages	2	2	1.00	1.00	Excellent
	3.2 Teacher selects the dominant student ability level to teach the curriculum 3.3 Teacher selects which skills and concepts to teach and which to emphasize to all groups of students	In your view to what extent do the following limit how you teach your math class? •students with different academic abilities •students who come from a wide range of backgrounds •students with special needs •uninterested students •disruptive students	1 = Not at all 2 = A little 3 = Some 4 = A lot	2	1	0.50	0.00	Poor
4.0 Quality of instructional delivery	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.2 Teacher uses teaching practices (coherent lessons) to produce students' academic achievement	Do you use a textbook in teaching math to your class?	1 = No 2 = Yes	2	1	0.50	0.00	Poor
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	Are the students in the TIMSS class permitted to use calculators during math lessons?	1 = unrestricted use 2 = restricted use 3 = calculators are not permitted	2	1	0.50	0.00	Poor

[a] I-CVI, item-level content validity index.

[b] k* = kappa designating agreement on relevance.

[c] Evaluation criteria for kappa, using guidelines described in Polit, Beck, & Owen (2007): Fair = k of .40 to .59; Good = k of .60–.74; and Excellent = k > .74.

Table 14 (continued)

Construct	Indicators	Questions	Scale	Number of Experts	Number Giving Rating of 3 or 4	I-CVI[a]	k*[b]	Evaluation[c]
4.0 Quality of instructional delivery	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	How often do you usually assign math homework?	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson	2	2	1.00	1.00	Excellent
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	If you assign math homework, how often do you assign each of the following tasks? •problem/question sets in textbook •small investigation or gathering data •finding one or more uses of the content covered	1 = Never 2 = Sometimes 3 = Always	2	2	1.00	1.00	Excellent
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	If you assign math homework, how often do you do the following tasks? •record whether or not the homework was completed •have students correct their own assignments in class •use it as a basis for class discussion •use it to contribute towards students' grades or marks	1 = Never 2 = Sometimes 3 = Always	2	2	1.00	1.00	Excellent
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.3 Teacher has a cognitive demand of the subject matter	How often do you interact with other teachers to discuss about how to teach a particular concept?	1 = Never or almost never 2 = 1-3 times a month 3 = 1-3 times a week 4 = Almost everyday	2	2	1.00	1.00	Excellent

[a] I-CVI, item-level content validity index.

[b] k* = kappa designating agreement on relevance.

[c] Evaluation criteria for kappa, using guidelines described in Polit, Beck, & Owen (2007): Fair = k of .40 to .59; Good = k of .60-.74; and Excellent = k > .74.

Table 14 (continued)

Construct	Indicators	Questions	Scale	Number of Experts	Number Giving Rating of 3 or 4	I-CVI[a]	k*[b]	Evaluation[c]
4.0 Quality of instructional delivery	4.3 Teacher has a cognitive demand of the subject matter	How well prepared do you feel you are to teach the following topics? •perimeter, area, and volume •coordinate geometry •algebraic representation •solving linear equations and inequalities •simple probabilities-understanding and calculations	1 = Not well prepared 2 = Somewhat prepared 3 = Very well prepared	2	2	1.00	1.00	Excellent
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	In your math lessons, how often do you usually ask students to do the following: •explain the reasoning behind an idea •represent and analyze relationships using tables, charts, or graphs •work on problems for which there is no immediately obvious method of solution •use computers to solve exercises or problems •write equations to represent relationships •practice computational skills	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson	2	2	1.00	1.00	Excellent
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.2 Teacher uses teaching practices (coherent lessons) to produce students' academic achievement	How often do students in your math class use calculators for the following activities: •checking answers •routine computation •solving complex problems •exploring number concepts	1= Never 2 = Some lessons 3 = Most lessons 4 = Every lesson	2	2	1.00	1.00	Excellent

[a] I-CVI, item-level content validity index.
[b] k* = kappa designating agreement on relevance.
[c] Evaluation criteria for kappa, using guidelines described in Polit, Beck, & Owen (2007): Fair = k of .40 to .59; Good = k of .60–.74; and Excellent = k > .74.

Table 15

TIMSS Teacher Survey Items Retained After Content Validation by OTL Indicator

Construct	Indicators	Questions	Scale
1.0 Content coverage (Total Items = 20)	1.1 Teacher arranges for all students to have access to the core curriculum 1.2 Teacher arranges for all students to have access to critical subject matter 1.3 Teacher ensures that there is curriculum content and test content overlap	The following list includes the main topics address by the TIMSS math test. Check the response that describe when students in your math class have been taught each topic. ***Major topics: Number, Algebra, Geometry, Measurement, Data and Chance***	1 = Taught before this year 2 = Mostly taught this year 3 = Not yet taught
2.0 Content exposure (Total Items = 9)	2.1 The amount of time teachers allocated to covering the content 2.3 Time devoted to a subject area (i.e. math)	How many minutes per week do you teach math to your math class?	Continuous
	2.2 Time allotted to students to learn 2.3 Time devoted to a subject area (i.e. math)	If you assign math homework, how many minutes of math homework do you usually assign your students?	1 = Less than 15 minutes 2 = 15-30 minutes 3 = 31-60 minutes 4 = 61-90 minutes 5 = More than 90 minutes
	2.1 The amount of time teachers allocated to covering the content 2.2 Time allotted to students to learn	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on each of the following activities? •homework review •lecture-style presentation by teacher •teacher-guided student practice •re-teaching and clarification of content/procedures •student independent practice •tests and quizzes •other	Percentages
3.0 Content emphasis (Total Items = 1)	3.1 Teacher selects topics within the curriculum to teacher (i.e. as a major topic, a minor review or not taught at all)	What subject matter do you emphasize the MOST in your math class?	1 = Number 2 = Algebra 3 = Geometry 4 = Combined Algebra and Geometry 5 = Combined Algebra, Geometry, Number 6 = Other

Table 15 (continued)

Construct	Indicators	Questions	Scale
4.0 Quality of instructional delivery (Total Items = 24)	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.2 Teachers uses teaching practices (coherent lessons) to produce students' academic achievement	How often do students in your math class use calculators for the following activities: •checking answers •routine computation •solving complex problems •exploring number concepts	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	How often do you usually assign math homework?	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	If you assign math homework, how often do you assign each of the following tasks? •problem/question sets in textbook •small investigation or gathering data •finding one or more uses of the content covered	1 = Never 2 = Sometimes 3 = Always
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	If you assign math homework, how often do you do the following tasks? •record whether or not the homework was completed •have students correct their own assignments in class •use it as a basis for class discussion •use it to contribute towards students' grades or marks	1 = Never 2 = Sometimes 3 = Always
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students 4.3 Teacher has a cognitive demand of the subject matter	How often do you interact with other teachers to discuss about how to teach a particular concept?	1 = Never 2 = 1-3 times a month 3 = 1-3 times a week 4 = Almost everyday
	4.1 Teacher uses varied teaching strategies and practices to meet the educational needs of all students	In your math lessons, how often do you usually ask students to do the following: •explain the reasoning behind an idea •represent and analyze relationships using tables, charts, or graphs •work on problems for which there is no immediately obvious method of solution •use computers to solve exercises or problems •write equations to represent relationships •practice computational skills	1 = Never 2 = Some lessons 3 = Most lessons 4 = Every lesson

Table 15 (continued)

Construct	Indicators	Questions	Scale
4.0 Quality of instructional delivery (Total Items = 24)	4.3 Teacher has a cognitive demand of the subject matter	How well prepared do you feel you are to teach the following topics? •perimeter, area, and volume •coordinate geometry •algebraic representation •solving linear equations and inequalities •simple probabilities-understanding and calculations	1 = Not well prepared 2 = Somewhat prepared 3 = Very well prepared

Empirical validation results. I performed an exploratory factor analysis (EFA) using Principal Components Analysis (PCA) to obtain evidence of and verify the internal structure of the theoretically-specified OTL measures, with 9 items from *content exposure* and 24 items *quality of instructional delivery* using three separate data sets—Data Set 1 (TIMSS 1999; $N_{\text{Data Set 1}} = 144$), Data Set 2 (TIMSS 2003; $N_{\text{Data Set 2}} = 146$), and Data Set 3 (TIMSS 2007; $N_{\text{Data Set 3}} = 215$). I assessed the other two dimensions in Stevens' OTL framework—*content coverage* and *content emphasis*—separately based on a descriptive data analysis of mathematics topics included in the TIMSS teacher surveys.

For the PCA, all items salient to a component were interpreted as a scale or dimension of OTL and interpreted according to Stevens' framework and the domain specified in Table 15. If the item sets were not validated across at least two data sets with item-component loadings more than |.32|, the items were dropped.

The initial EFA extracted 10 components from Data Set 1 and 11 components for Data Set 2 and Data Set 3, accounting for 62%, 68%, and 65% of total variance, respectively. However, the scree plots suggested six components for all three data sets. My examination rendered 17 items that loaded on two theoretically-aligned OTL components: 14 items from Components 1, 2, 3, and 5 loaded on the *quality of instructional delivery* OTL component (with robust loadings of 0.32-0.88). However, just three items from Components 4 and 6 loaded on the *content exposure* OTL component (with loadings of 0.41-0.78), suggesting difficulties in scaling that construct using the TIMSS teacher survey items. Table 16 presents the loading and the variance extracted from the PCA.

Following EFA, I assessed convergent validity with the Pearson product-moment correlations of total scale scores from items in the two derived components. The correlations between the *quality of instructional delivery* OTL and *content exposure* OTL components were: 0.06 (Data Set 1), -0.02 (Data Set 2), and 0.05 (Data Set 3). These correlations were unacceptable vis-à-vis the expected correlations of 0.30-0.50 for theoretically-specified OTL measures (Herman et al., 2000).

I then estimated the scale score reliability levels. The reliability coefficients ranged from 0.70 to 0.79 across the three data sets for the *quality of instructional delivery* OTL component. However, the reliability coefficients ranged from -0.35 to 0.10 across the three data sets for the *content exposure* OTL component. Considering the extremely low reliability, I decided to delete the *content exposure* OTL component as a defensible scale.

Though I was unable to validate a multi-item scaled measure of the *content exposure* OTL (i.e., amount of time teachers allocated to covering the content), I decided to keep one TIMSS teacher survey item from this subdomain for the next quantitative phase: *How many minutes per week do you teach math to your math class?* This single-item variable served as a proxy for *content exposure* OTL and is referred to as

Table 16

Principal Component Analysis (PCA) of TIMSS 1999, 2003, and 2007: Loading Estimates, Variance Extracted, and Reliability

		TIMSS 1999 ($N_{\text{Data Set 1}}$=144)						TIMSS 2003 ($N_{\text{Data Set 2}}$=146)						TIMSS 2007 ($N_{\text{Data Set 3}}$=215)					
Item	Description	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	How well prepared do you feel you are to teach solving linear equations and inequalities?	**0.88**						**0.83**						**0.81**					
2	How well prepared do you feel you are to teach algebraic representation?	**0.82**						**0.81**						**0.68**					
3	How well prepared do you feel you are to teach understanding and calculating simple probabilities?	**0.73**						**0.81**						**0.66**					
4	How well prepared do you feel you are to teach coordinate geometry?	**0.70**						**0.85**						**0.86**					
5	How well prepared do you feel you are to teach perimeter, area, and volume?	**0.65**						**0.80**						**0.83**					
6	In your math lessons, how often do you ask the students to represent and analyze relationships using tables, charts, or graphs?		**0.68**						**0.76**						**0.66**				
7	In your math lessons, how often do you ask the students to explain the reasoning behind an idea?		**0.66**						**0.40**						**0.63**				
8	In your math lessons, how often do you ask the students to write equations to represent relationships?		**0.64**						**0.78**						**0.74**				
9	In your math lessons, how often do you ask the students to work on problems for which there is no immediate solution?		**0.59**						**0.60**						**0.53**				
10	If you assign math homework, how often do you have students correct their own assignments in class?		0.43									**0.78**							**0.45**

Notes: Varimax-rotated PCA was conducted. Loadings of retained components are presented in boldface fonts. Components 1-3 and 5 included items matching *quality of instructional delivery* OTL. Components 4 and 6 included items matching *content exposure* OTL.

Table 16 (continued)

		TIMSS 1999 ($N_{\text{Data Set 1}}$=144)						TIMSS 2003 ($N_{\text{Data Set 2}}$=146)						TIMSS 2007 ($N_{\text{Data Set 3}}$=215)					
Item	Description	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
11	If you assign math homework, how often do you use it to contribute to students' marks?			**0.76**									**0.67**				**0.64**		
12	If you assign math homework, how often do you record whether or not homework was completed?			**0.74**							**0.75**						**0.75**		
13	In your math lessons, how often do you ask the students to practice computational skills?			**0.49**														**0.58**	
14	How often do you usually assign math homework?				**0.65**					**0.84**						**0.88**			
15	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on independent practice?				-0.64							**0.44**							**0.41**
16	If you assign math homework, how often do you have students correct their own assignments in class?				**0.47**						**0.78**						**0.47**		
17	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on re-teaching and clarification of content/procedures?				-0.42						0.41								
18	If you assign math homework, how many minutes of math homework do you usually assign to your students?					**0.64**							**0.78**				**0.47**		
19	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on tests and quizzes?					**0.63**				**0.37**									
20	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on homework reviews?					**0.61**									**0.57**			**-0.32**	

Notes: Varimax-rotated PCA was conducted. Loadings of retained components are presented in boldface fonts. Components 1-3 and 5 included items matching *quality of instructional delivery* OTL. Components 4 and 6 included items matching *content exposure* OTL.

Table 16 (continued)

		TIMSS 1999 ($N_{\text{Data Set 1}}$=144)						TIMSS 2003 ($N_{\text{Data Set 2}}$=146)						TIMSS 2007 ($N_{\text{Data Set 3}}$=215)					
Item	Description	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
21	If you assign math homework, how often do you use it as a basis for class discussion?								**0.47**										
22	If you assign math homework, how often do you assign problem/question sets in textbook?									**0.86**						**0.81**			
23	In a typical week of math lessons for the TIMSS class, what percentage of time do students spend on teacher-guided student practice?									-0.43									
24	If you assign math homework, how often do you assign finding one or more uses of the content covered?											**0.69**					-0.32	**0.32**	
25	If you assign math homework, how often do you assign small investigation or gathering data?											0.79							
26	How often do you interact with other teachers to discuss about how to teach a particular concept?											**0.35**						**0.76**	
27	How many minutes per week do you teach math to your math class?																	0.36	
	Variance extracted	11%	21%	27%	33%	38%	43%	12%	20%	27%	33%	38%	44%	11%	19%	25%	31%	36%	41%

Notes: Varimax-rotated PCA was conducted. Loadings of retained components are presented in boldface fonts. Components 1-3 and 5 included items matching *quality of instructional delivery* OTL. Components 4 and 6 included items matching *content exposure* OTL.

instructional time hereinafter. This decision was motivated by the vast literature documenting the positive relationship between instructional time and student achievement (Carroll, 1963; Carnoy, Khavenson, Loyalka, Schmidt, & Zakharov, 2016; Floden, 2002; Husén, 1967; Kifer & Burstein, 1992; Schmidt & Burroughs, 2013; Schmidt & Maier, 2009; Schmidt, Zoido, & Cogan, 2013; Travers & Westbury, 1989; Wiley & Harnischfeger, 1974). The limitations of this researcher's decision will be acknowledged in the discussion section.

Research Question Five results. Results of how OTL changed in the pre-*yutori*, mid-*yutori*, and post-*yutori* stages are presented below by OTL measure per Stevens' framework: *content coverage, content emphasis, instructional time,* and *quality of instructional delivery*. As indicated previously, the changes in *content coverage* and *content emphasis* over time were examined using a descriptive statistical analysis of the TIMSS teacher surveys (frequencies indicating whether a topic was taught and percents of students exposed); changes in *instructional time* (in minutes per week) and *quality of instructional delivery* (total scale score based on validation results in Tables 14-15) were examined using a one-way ANOVA.

Changes in content coverage. Table 12 (presented previously in this chapter) presented the topics covered in the implemented 8th grade mathematics curriculum and the teachers' reports about the percent of students taught the topics. Looking across the three rounds of the TIMSS, teachers taught almost equal numbers of fraction and number sense-related topics, measurement-related topics, and algebra-related topics to similar percentages of students. Conversely, more variability was observed in teachers' self-reported coverage of the topics from data representation, analysis, probability, and geometry. For example, teachers reported teaching “representation and interpretation of data in graphs, charts, and tables” to more students in 2003 (55%) and 2007 (52%) than in 1999 (43%). Similar discrepancies were observed for “simple probabilities—understanding and calculations” and “organizing a set of data by one or more characteristics using a tally chart, table, or graph” topic. For geometry, teachers reported teaching “symmetry and transformations” to 88% of students in 2003, which was much lower than percentages of students taught in 1999 (98%) and 2007 (99%). Another big gap was observed for the “relationship between 2-dimensional and 3-dimensional shapes”: this topic was taught to 51% of students in 2003 but to 89% of students in 2007. In general, I did not see a consistent pattern in content coverage over time.

Numerous gaps between the intended curriculum and the implemented curriculum were observed across three rounds of the TIMSS. For example, at least 90% of Japanese 8th graders were expected to be taught “representation and interpretation of data in graphs, charts, and tables” in 1999, 2003, and 2007 (as indicated by a solid circle in Table 12). But only 43%, 55%, and 52% of 8th graders actually received instruction on that topic, as reported by their teachers. Similar gaps were observed for “simple probabilities—understanding and calculations,” “organizing a set of data by one more characteristics using a tally chart, table, or graph,” “relationship between two-dimensional and three-dimensional shapes,” and “writing expressions for general terms in number pattern sequence.”

Interestingly, I also found evidence of teachers not following the intended curriculum, as prescribed by the Japanese government. For example, “translation, reflection, rotation, and enlargement” was intended to be taught in Grade 9 (as indicated by an empty circle in Table 12), but over half of the students learned this topic in 8th grade in 2003 and 2007. Similarly, “estimates of measurement, accuracy of measurement” was a topic intended to be taught in Grades 10-12, according to the Japanese NRC surveyed in 2003. But teachers reported teaching this topic to 43% of their students.

Changes in content emphasis. As shown previously in Table 14, only a single item was retained for the *content emphasis OTL* subdomain after content validation. This TIMSS teacher survey item asked: *What subject matter do you emphasize the MOST in your math class?* Teachers were asked to select only one choice out of six response options, including (a) number; (b) algebra; (c) geometry; (d) combined algebra and geometry; (e) combined algebra, geometry, number; and (f) other.

Using frequency counts, I found that “combined algebra and geometry” was the emphasis identified by most teachers in 8th grade mathematics classes, as reported by teachers surveyed in TIMSS 1999, TIMSS 2003, and TIMSS 2007. Furthermore, I found similar percentages of teachers reporting an emphasis on this topic combination: 35% in 1999, 34% in 2003, and 33% in 2007. This implied that there was no change in the *content emphasis OTL* measure in 8th grade mathematics class over time, according to teacher self-reports.

Changes in instructional time. As mentioned, *instructional time* was a proxy for *content exposure.* Because the homogeneity of variance assumption was not met, I performed a Welch ANOVA with Games-Howell post-hoc test to compare mean difference in *instructional time* across three consecutive rounds of the TIMSS. As shown in Table 17, significant differences were found in *instructional time*, $F(2,448) = 205.23$, p

< .001. Games-Howell post-hoc tests revealed that teachers reported significantly longer *instructional time* for their 8th grade mathematics class in TIMSS 1999 (M = 197.97, SD = 17.23) than in TIMSS 2003 (M = 157.47, SD = 32.04) and in TIMSS 2007 (M = 156.62, SD = 21.28). This suggested that the weekly instructional time for 8th grade mathematics classes decreased by at least 20% or by 40 minutes per week after the implementation of *yutori* reforms. In other words, *yutori* reforms contributed to a 20% drop in *instructional time* in 8th grade mathematics class.

Changes in quality of instructional delivery. Table 18 presents the mean comparison in the validated *quality of instructional delivery* OTL measure over time. Using the pooled data set, I first created scale scores for the *quality of instructional delivery* OTL by summing the standardized scores of all 14 items on the scale. Then I checked the assumptions for a one-way ANOVA prior to conducting the analysis. Since all assumptions were satisfied, I conducted a one-way ANOVA with Tukey HSD post-hoc tests. Significant differences were observed in teachers' perception of readiness to teach OTL measure, F(2,367) = 18.35, p < .001. Post-hoc Tukey HSD tests revealed that teachers reported a significantly higher level of the *quality of instructional delivery* OTL in the TIMSS 2007 (M = 2.33, SD = 5.40) than in the TIMSS 2003 (M = -0.17, SD = 5.81) and in the TIMSS 1999 (M = -2.37, SD = 7.05). To interpret the difference in the *quality of instructional delivery* OTL, I presented teachers' responses to items on their original scoring scale in Table 19. Proportionally, more post-*yutori* teachers reported being prepared to teach mathematics topics (i.e., perimeter, area, volume, geometry, and so on) than pre-*yutori* and mid-*yutori* teachers. After *yutori* reforms were implemented, teachers reported more of the following homework-related strategies: assigning problem sets in textbook as math homework, having students correct their own assignments in class, and using math homework to contribute to students' marks.

In terms of *quality of instructional delivery*, the post-*yutori* cohort teachers were found to be the most prepared to teach mathematics and utilize the most teaching strategies to meet their students' educational needs, followed by the mid-*yutori* cohort teachers, and lastly, the pre-*yutori* cohort teachers.

Table 17

Mean Comparison in Instructional Time Measure Between TIMSS 1999, TIMSS 2003, and TIMSS 2007

	N	Mean	Standard Deviation	F	*p*
TIMSS 1999	138	197.97	17.23		
TIMSS 2003	146	157.47	32.04	205.23	<0.001[a, b, c]
TIMSS 2007	167	156.62	21.28		

Notes: Weekly instructional time is given in minutes.
Homogeneity of variance assumption was not met.
Welch ANOVA with Games-Howell post-hoc test was conducted.
[a] Significant difference between TIMSS 1999 and TIMSS 2003, [b] significant difference between TIMSS 1999 and TIMSS 2007, [c] significant difference between TIMSS 2003 and TIMSS 2007.

Table 18

Mean Comparison in Quality of Instructional Delivery OTL Measure Between TIMSS 1999, TIMSS 2003, and TIMSS 2007

	N	Mean	Standard Deviation	F	*p*
TIMSS 1999	96	-2.37	7.05		
TIMSS 2003	126	-0.17	5.81	18.35	<0.001[a, b, c]
TIMSS 2007	148	2.33	5.40		

Notes: Standardized *quality of instructional delivery* OTL scale scores were analyzed. One-way ANOVA with Tukey HSD post-hoc test was performed.
Mean, SD, and range for the pooled raw scale scores on *Quality of Instructional Delivery* are: M=33.55, SD=4.39, Range: (21.00-46.00).
[a] Significant difference between TIMSS 1999 and TIMSS 2003, [b] significant difference between TIMSS 1999 and TIMSS 2007, [c] significant difference between TIMSS 2003 and TIMSS 2007.

Table 19

Teacher Responses to Quality of Instructional Delivery OTL Measures on TIMSS 1999, TIMSS 2003, and TIMSS 2007

Questions	Response Options	TIMSS 1999 (*n*=96)		TIMSS 2003 (*n*=126)		TIMSS 2007 (*n*=148)	
		Frequency	%	Frequency	%	Frequency	%
How often do you usually assign math homework?	Never or almost never	13	9%	0	0%	0	0%
	Some lessons	**104**	**72%**	**80**	**60%**	**91**	**43%**
	Most lessons	18	13%	23	17%	61	29%
	Every lesson	9	6%	30	23%	58	28%
If you assign math homework, how often do you assign problem/question sets in textbook?	Never or almost never	**70**	**53%**	1	1%	0	0%
	Sometimes	47	36%	**78**	**59%**	102	49%
	Always or almost always	14	11%	54	41%	**107**	**51%**
If you assign math homework, how often do you record whether or not homework was completed?	Never or almost never	23	22%	4	3%	7	3%
	Sometimes	31	29%	54	41%	64	30%
	Always or almost always	**52**	**49%**	**75**	**56%**	**140**	**66%**
If you assign math homework, how often do you have students correct their own assignments in class?	Never or almost never	**46**	**43%**	39	29%	46	22%
	Sometimes	38	36%	**71**	**53%**	**108**	**51%**
	Always or almost always	22	21%	23	17%	56	27%
If you assign math homework, how often do you use it to contribute to students' marks?	Never or almost never	**64**	**60%**	34	26%	72	34%
	Sometimes	21	20%	**64**	**48%**	**100**	**47%**
	Always or almost always	21	20%	35	26%	39	19%
In your math lessons, how often do you ask the students to work on problems for which there is no immediate solution?	Never or almost never	7	5%	2	1%	15	7%
	Some lessons	**76**	**54%**	**83**	**58%**	**144**	**67%**
	Most lessons	39	28%	44	31%	41	19%
	Every lesson	20	14%	15	10%	14	7%
In your math lessons, how often do you ask the students to write equations to represent relationships?	Never or almost never	0	0%	0	0%	1	1%
	Some lessons	28	19%	57	39%	78	36%
	Most lessons	**80**	**56%**	**70**	**48%**	**94**	**44%**
	Every lesson	36	25%	18	12%	41	19%
How well prepared do you feel you are to teach perimeter, area, and volume?	Not well prepared	18	13%	8	6%	4	2%
	Somewhat prepared	**86**	**61%**	**95**	**67%**	67	32%
	Very well prepared	37	26%	38	27%	**141**	**67%**
How well prepared do you feel you are to teach coordinate geometry?	Not well prepared	25	18%	3	2%	2	1%
	Somewhat prepared	**79**	**57%**	**81**	**57%**	72	34%
	Very well prepared	34	25%	57	40%	**137**	**65%**
How well prepared do you feel you are to teach algebraic representation?	Not well prepared	14	10%	10	7%	11	5%
	Somewhat prepared	**87**	**60%**	**91**	**65%**	**105**	**51%**
	Very well prepared	43	30%	40	28%	92	44%
How well prepared do you feel you are to teach solving linear equations and inequalities?	Not well prepared	10	7%	1	1%	1	1%
	Somewhat prepared	**79**	**55%**	**80**	**57%**	59	28%
	Very well prepared	55	38%	60	43%	**148**	**71%**
How well prepared do you feel you are to teach understanding and calculating simple probabilities?	Not well prepared	36	25%	13	9%	12	6%
	Somewhat prepared	**78**	**55%**	**82**	**58%**	96	47%
	Very well prepared	28	20%	46	33%	**97**	**47%**

Notes: The most frequent response choices are presented in boldface fonts.

Summary of Key Findings

From the researcher interviews, I found that *yutori* reforms were not implemented as MEXT had intended because there were several individual teacher-level barriers and institutional-level barriers to implementation. Teachers and schools, being entrusted with much autonomy, followed their ideals to prepare students for higher achievement on high-stakes examinations at the next level of schooling and subverted the implementation of *yutori* policies. The sole facilitator of *yutori* reforms, pilot schools, did not have much influence on the implementation process as the pilot school programs were only available to a few designated teachers at each school. Another observation was that students and parents relied on *juku* to fill the gaps in knowledge stemming from the cuts in instructional time in the core academic subjects. However, the *jukus* imposed heavy financial burdens on families. Such heavy financial burdens may have widened educational inequalities between the students from low socioeconomic levels who could not afford the *jukus* and those from higher socioeconomic backgrounds.

The *quality of instructional delivery* scale of Stevens' multidimensional framework was validated with the TIMSS 1999, 2003, and 2007 survey data from samples of participating Japanese 8th grade mathematics teachers. Two other OTL measures—*content coverage* and *content emphasis*—were retained after content validation, but could not be empirically validated as scales because they yielded disparate item-level distributions as frequency counts on various topics. Along with *quality of instructional delivery* OTL, I also kept a single-item variable, *instructional time*, going into the next quantitative analysis phase of this study.

Lastly, I compared the changes in OTL over time in 8th grade mathematics classes, as given by the TIMSS teacher survey data from 1999, 2003, and 2007. In terms of *content coverage*, I found some discrepancies in teachers' self-reported coverage of the topics from data representation, analysis, probability, and geometry, but not in fraction and number sense-related topics, measurement-related topics, and algebra-related topics.

Further, I also observed two types of gaps between the intended curriculum reported by NRC surveys and the implemented curriculum as reported by teachers. One gap was that teachers did not teach the topics outlined in the intended curriculum to their students. In other words, teachers did not fulfill MEXT's curricular coverage requirements. Another kind of gap was that teachers explicitly taught mathematics topics that were intended as preparation for higher grade levels. This observation was corroborated by my interview-based finding that teachers continued to cover the content of the pre-*yutori* curriculum even after the official implementation of *yutori* reforms began.

With regard to *content emphasis*, I found that most teachers identified an emphasis on "combined algebra and geometry" in the 8th grade mathematics classes in TIMSS 1999, TIMSS 2003, and TIMSS 2007.

I found that weekly *instructional time* for 8th grade mathematics classes decreased by at least 20% or by 40 minutes per week after the implementation of *yutori* reforms. In other words, *yutori* reforms contributed to a 20% drop in *instructional time* in 8th grade mathematics class, based on the TIMSS survey data. This suggested a structural match with the intent of the *yutori* reforms.

Lastly, in terms of *quality of instructional delivery*, I found that the post-*yutori* cohort of teachers were the most prepared to teach mathematics based on survey item indicators that matched Stevens' framework, and utilized the most varied teaching strategies to meet their students' educational needs, followed by the mid-*yutori* cohort teachers, and lastly, the pre-*yutori* teacher cohorts.

Recapitulation of Research Question, Analysis Methods, and Data Sources

Research Question Six asked: To what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement: (a) at the classroom level, to what extent did the observed changes in OTL levels over time affect changes in aggregated 8th grade students' mathematics achievement between cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori*)? (b) using a multilevel approach, to what extent did the observed changes in OTL levels over time affect changes in 8th grade students' mathematics achievement within cohorts? (c) did OTL moderate the relationship between students' mathematics achievement and socioeconomic background within cohorts, as measured by the TIMSS student assessments?

To address this research question, I conducted quantitative analyses in three stages. First, I analyzed the OTL-student achievement relationship at the classroom level between the three cohorts (pre-*yutori*, mid-*yutori*, and post-*yutori*) using analysis of covariance (ANCOVA). Examining the effects of OTL on aggregated student achievement at classroom level was logical for this study because OTL is a teacher-level variable, per Stevens' (1996) OTL framework. Also, because of the sampling design, the participating students were nested observations because TIMSS samples students from intact classrooms. Using aggregated student achievement in mathematics as an outcome variable and *yutori* cohort by year and OTL as independent variables, I analyzed three

ANCOVA models: Model 1 used *instructional time* as the independent variables; Model 2 used *quality of instructional delivery* OTL as the independent variables; Model 3 used both *instructional time* and *quality of instructional delivery* OTL as independent variables. To make a meaningful interpretation of the OTL-achievement relationship, I also ran an ANOVA model using aggregated student achievement in mathematics as the outcome variable and *yutori* cohort by year as the independent variable. The ANOVA model served as a baseline against which ANCOVA models were compared.

In the second stage, I examined the association between classroom-level OTL and student-level mathematics achievement using hierarchical linear models (HLM). Because of the nested data structure of students in classrooms/schools, the students were not statistically independent observations. HLM accounted for the within-classroom dependencies and allowed for the investigation of the OTL at multiple levels.

The dependent variable in HLM analyses was the TIMSS 8th grade mathematics achievement score, which consisted of not one but rather five plausible values given by the TIMSS. As indicated in Chapter III, five plausible values were randomly drawn from the conditional distribution of proficiency scores for each student (Gonzalez & Miles, 2001). Each plausible value provided information about each student's proficiency level as well as information about the uncertainty in the score. I used HLM 7 software for HLM modeling because HLM 7 can perform analyses on all TIMSS plausible values for mathematics and has the ability to accommodate survey weights performance (von Davier, Gonzalez, & Mislevy, 2009). As described in Chapter III, the HLM analyses were conducted for the two OTL measures—*instructional time* and *quality of instructional delivery*—within cohorts in five steps.

Lastly, I added a cross-level interaction between student SES level and OTL (defined as *instructional time*) to see if OTL moderated the relationship between students' mathematics achievement and socioeconomic background within cohorts. This model would speak to potential educational inequalities by revealing whether OTL in terms of greater instructional time mitigated or exacerbated the relationship between students' mathematics achievement and socioeconomic background within cohorts.

Overall Results in Tables and Descriptions

Classroom-level analysis of OTL results. Table 20 presents the mean comparisons in aggregated mathematics achievement at the classroom level between cohorts. I found significant differences in mean mathematics achievement between cohorts by year, F(2,298.38) = 3.73, p = 0.025. The Games-Howell post-hoc test showed that the pre-*yutori* cohort (as indicated by the TIMSS 1999) had significantly higher (p = 0.026) aggregated mean mathematics achievement than the mid-*yutori* cohort (as indicated by the TIMSS 2003). There were no significant differences between the pre-*yutori* cohort and the post-*yutori* cohort (as indicated by the TIMSS 2007). This between-cohort comparison served as a baseline to which I compared the ANCOVA results.

Tables 21-23 present the ANCOVA results. There was a significant difference in mean classroom-level mathematics achievement [F(2, 366) = 4.06, *p* = 0.018] between cohorts, while adjusting for *quality of instructional delivery* OTL levels (see Table 21). Holding the *quality of instructional delivery* OTL level constant, a Bonferroni post-hoc test found that the pre-*yutori* cohort had significantly higher aggregated mathematics achievement than the mid-*yutori* cohort and the TIMSS 2003 (*p* = 0.023) and the post-

yutori cohort (p = 0.054). This confirmed my hypothesis that *instructional time* changes under *yutori* possibly resulted in these changes in outcomes.

Controlling for *instructional time*, I did not find significant differences in mean classroom-level mathematics achievement [F(2, 447) = 0.37, p = 0.690] between cohorts (see Table 22). But, as illustrated in Table 23, a similar significant effect was found for both *instructional time* and *quality of instructional delivery* OTL levels on achievement, but there was no significant difference in mean classroom-level mathematics achievement by year [F(2, 364) = 0.37, p = 0.694]. Together, the results of the three ANCOVAs suggested the potency of the *instructional time* component of OTL over the *quality of instructional delivery* OTL in affecting the students' achievement levels. At the same time, both OTL construct measures, *instructional time* and *quality of instructional delivery*, had statistically significant effects on students' mathematics achievement as independent variables, although year of reforms did not (see Table 22). These findings suggested that *quality of instructional delivery* played a significant role in affecting mean classroom-level mathematics achievement, even after controlling for *instructional time* and years of reform.

Interpreting the ANCOVA results vis-à-vis the baseline ANOVA results, I inferred that the class-level mathematics achievement gaps across time were correlated with the cut in *instructional time*, as outlined in the *yutori* reforms. This between-cohort achievement gap would disappear if all three cohorts had been given equal amounts of weekly mathematics instructional time.

Another interesting finding was that the post-*yutori* teachers were able to narrow the achievement gap between them and the pre-*yutori* cohort by increasing their *quality* *of instructional delivery* OTL levels. This finding was suggested by the non-significant difference between the pre-*yutori* and the post-*yutori* cohorts in the baseline ANOVA model (see Table 20) and the significant post-hoc test result between the pre-*yutori* and the post-*yutori* cohorts in the ANCOVA model that controlled for *quality of instructional delivery* (see Table 21). In other words, this finding suggested that the post-*yutori* cohort would have significantly lower mathematics achievement than the pre-*yutori* cohort if both cohorts had been given the same level of *quality of instructional delivery* OTL in their mathematics classes.

Table 20

ANOVA Mean Comparison of Aggregated Mathematics Achievement at Classroom Level Between TIMSS 1999, TIMSS 2003, and TIMSS 2007

Year	N	Mean	SD	F	*P*
TIMSS 1999	140	576.41	1.47	3.73	0.025[a]
TIMSS 2003	146	567.64	1.78		
TIMSS 2007	169	570.09	1.44		

Notes: Homogeneity of variance assumption was not met. Welch ANOVA with Games-Howell post-hoc test was conducted. [a] Games-Howell post-hoc test revealed significant difference between TIMSS 1999 and TIMSS 2003 (p = 0.026).

Table 21

ANCOVA Summary Controlling for Quality of Instructional Delivery OTL

Source	Sum of Squares	*df*	Mean Square	F	*p*
Quality of Instructional Delivery OTL	13035.13	1	13035.13	11.18	0.001
Year	9461.11	2	4730.56	4.06	0.018 [a, b]
Error	426574.07	366	1165.50		

Notes: [a] Bonferroni pairwise comparisons revealed significant difference between TIMSS 1999 and TIMSS 2003 (p = 0.023); [b] Bonferroni pairwise comparisons revealed borderline significant difference between TIMSS 1999 and TIMSS 2007 (p = 0.054).

Table 22

ANCOVA Summary Controlling for Instructional Time

Source	Sum of Squares	*df*	Mean Square	F	*p*
Instructional Time	7674.97	1	7674.97	6.93	0.009
Year	821.62	2	410.81	0.37	0.690
Error	495213.86	447	1107.86		

Table 23

ANCOVA Summary Controlling for Quality of Instructional Delivery OTL and Instructional Time

Source	Sum of Squares	*df*	Mean Square	F	*p*
Instructional Time	9469.97	1	9469.97	8.34	0.004
Quality of Instructional Delivery OTL	13249.44	1	13249.44	11.66	0.001
Year	832.36	2	416.18	0.37	0.694
Error	413525.34	364	1136.05		

Multilevel analysis of OTL results. Table 24-26 present the within-cohort HLM analyses results for pre-*yutori* (TIMSS 1999), mid-*yutori* (TIMSS 2003), and post-*yutori* (TIMSS 2007), respectively. The unconditional model (Model A), which is a model without any independent variables as student or classroom predictors, was estimated first. In this unconditional model, the total variance in student mathematics achievement was decomposed into variation between classrooms (level 2) and variation between students within classrooms (level 1). The intraclass correlation coefficient (ICC) suggested that 8% (for pre-*yutori* cohort), 17% (for mid-*yutori* cohort), and 19% (for post-*yutori* cohort) of the variation in mathematics achievement were associated with differences between classrooms, respectively. The ICCs substantiated the application of HLM models for the mid-*yutori* cohort and the post-*yutori* cohort, but not for the pre-*yutori* cohort. Thus, I performed ordinary least squares (OLS) regression analyses instead of HLM to examine the TIMSS 1999 data. Multilevel analysis results are presented by cohort below.

OLS analysis of OTL results for pre-yutori cohort. Table 24 presents the OLS analyses results for pre-*yutori* cohort (TIMSS 1999). In Model 2, I added student-level and classroom-level covariates to the OLS model. Student SES ($b = 15.71$, $p < 0.001$) and class SES ($b = 21.36$, $p < 0.05$) were found to be strong, positive predictors of students' mathematics achievement, even after controlling for all other variables. This result was consistent with the literature (Berliner, 2006; Blanden, Gregg, & Machin, 2005; Bradley & Corwyn, 2002; Davis-Kean, 2005).

I also found that male students performed better on the TIMSS 1999 mathematics achievement than their female peers ($b = 6.75$, $p < 0.05$) and that student age was a significant predictor of mathematics achievement for the pre-*yutori* cohort (*b* = 22.23, p < 0.001).

In Model 3, I added the *quality of instructional delivery* OTL measure to the OLS model. Holding everything constant, there was no significant association between the *quality of instructional delivery* OTL and students' mathematics achievement on TIMSS 1999 (*p* > 0.05). But all the significant covariates in Model 2 remained significantly correlated with students' mathematics achievement in Model 3 as well.

In Model 4, I estimated the effects of *instructional time* on mathematics achievement while controlling for student- and teacher-level covariates. Holding everything constant, there was no significant association between *instructional time* and students' mathematics achievement on the TIMSS 1999 (*p* > 0.05). But all the significant covariates in Model 2 remained significantly correlated with students' mathematics achievement in Model 3 as well.

In Model 5, I added both the *instructional time* and *quality of instructional delivery* OTL measure to Model 2. Controlling for *instructional time* and all other variables, there was no statistically significant correlation between the *quality of instructional delivery* OTL measure and 8th graders' achievement on TIMSS 1999 (*p* > 0.05). Similarly, there was no statistically significant correlation between *instructional time* and 8th graders' achievement on TIMSS 1999 (*p* > 0.05) after holding the *quality of instructional delivery* OTL measure and all other variables constant. However, I found that the correlation between class SES and mathematics achievement failed to reach statistical significance (*p* > 0.05).

Table 24

Hierarchical Linear Models Using Stevens' OTL Measure to Predict 8th Grade Mathematics Achievement in TIMSS 1999 ($n_{unweighted}$ = 4606)

	Model 1	Model 2	Model 3	Model 4	Model 5
Student-level variables					
Female Student		-6.75*	-6.65*	-6.00*	-6.76*
		(3.00)	(3.26)	(2.79)	(3.33)
Student Age		22.23***	24.68***	23.86***	24.59***
		(4.88)	(5.35)	(4.22)	(5.30)
Student SES		15.71***	15.38***	15.50***	15.38***
		(1.57)	(1.93)	(1.58)	(1.93)
Teacher-level variables					
Intercept	579.60***	209.10***	183.46*	191.48**	187.05*
	(3.05)	(68.51)	(68.51)	(68.51)	(85.69)
Class SES		21.36*	20.70*	20.83*	20.81
		(9.61)	(13.32)	(9.98)	(13.51)
Female Teacher		-5.16	-7.66	-4.33	-7.85
		(3.93)	(4.38)	(3.78)	(4.38)
Teacher Age [a]					
Age 25-29		-8.68	-16.34	-8.93	-15.65
		(12.74)	(17.40)	(12.07)	(17.80)
Age 30-39		-2.44	-11.18	-4.82	-10.27
		(15.10)	(19.38)	(14.48)	(19.74)
Age 40-49		3.55	1.66	0.42	-0.62
		(17.50)	(22.17)	(17.66)	(23.06)
Age 50-59		2.83	-13.33	-7.15	-10.55
		(2.56)	(25.33)	(22.40)	(26.98)
Age 60 and over		16.03	0.73	9.98	2.46
		(26.14)	(33.29)	(26.39)	(35.07)
Teacher Experience		0.00	0.05	0.22	0.00
		(0.71)	(0.91)	(0.73)	(0.96)
Number of Students in Class		1.55	1.51	1.33	1.44
		(0.79)	(1.01)	(0.77)	(1.06)
Quality of Instructional Delivery OTL			0.53		0.54
			(0.38)		(0.38)
Instructional Time				0.81	1.14
				(2.71)	(2.61)
Random components					
Level-1 variance	5587.49				
Level-2 variance	493.71				
Intraclass correlation coefficient (ICC)	0.08				
% change in Level-2 variance compared to the null model	-				

Notes: Standard errors in parentheses. Student weights were applied. Models 2 to 5 were estimated with using OLS regression with standard robust errors. [a] Reference group for Teacher Age—Under age 25.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

HLM analysis of OTL results for mid-yutori cohort. Table 25 presents two-level HLM models estimating the relationship between OTL measures and 8th grade mathematics achievement for the mid-*yutori* cohort.

In Model 2, the addition of the student-level and teacher-level covariates explained 64% of the estimated achievement variance between classes. Student SES (b = 14.52, p < 0.001) and class SES (b = 63.90, p < 0.001) were found to be strong positive predictors of mathematics achievement, even after controlling for all other variables. Student age was also a significant predictor of 8th grade mathematics achievement for the mid-*yutori* cohort (b = 9.23, p < 0.05).

Model 3 estimated the relationship between *quality of instructional delivery* OTL measure and the students' mathematics performance on the TIMSS 2003, adjusted for student-level and classroom-level covariates. Holding everything constant, there was no significant association between the *quality of instructional delivery* OTL measure and the class-level mathematics performance (b = 0.42, p > 0.05). Using the *quality of instructional delivery* OTL measure as a predictor reduced between-class variance by less than 1%, when compared to Model 2.

Model 4 estimated the relationship between *instructional time* and the 8th graders' mathematics performance on the TIMSS 2003, adjusted for student-level and classroom-level covariates. There was a significant association between *instructional time* and the class-level mathematics performance on the TIMSS 2003 (b = 7.06, p < 0.001). Since the *instructional time* variable was standardized, this coefficient estimate indicated that for a 1 standard deviation unit increase in *instructional time* for a mathematics class, the class-level mathematics achievement on the TIMSS 2003 would be expected to increase by 7 scaled score points. The addition of *instructional time* further reduced the between-classroom variance in achievement by 5% when compared to Model 2.

In Model 5, I estimated the relationship between both the *instructional time* and *quality of instructional delivery* OTL measure and the 8th graders' mathematics performance on the TIMSS 2003. Though there was no significant relationship between the *quality of instructional delivery* OTL measure and mean mathematics achievement (b = 0.31, p > 0.05), there was a significant *instructional time* effect on the class-level mathematics assessment scores on the TIMSS 2003 with the effect equal to 6.94 (SE = 1.89, p < 0.001). In other words, the model predicted a 7-point gain on the TIMSS 2003 mathematics assessment for a student whose weekly mathematics class time increased by a 1 standard deviation unit, controlling for everything else. The addition of the *instructional time* and *quality of instructional delivery* OTL measure reduced the between-class variance by 71% when compared to the null model (Model 1).

HLM analysis of OTL results for post-yutori cohort. Table 26 presents two-level HLM models estimating the relationship between OTL measures and 8th grade mathematics achievement for the post-*yutori* cohort.

In Model 2, the addition of the student-level and teacher-level covariates explained 64% of the variation between class. Student SES (b = 20.40, p < 0.001) and class SES (b = 54.29, p < 0.001) were found to be strong positive predictors of mathematics achievement, even after controlling for all other variables. Student age was also a significant predictor of 8th grade mathematics achievement for the post-*yutori* cohort (b = 10.57, p < 0.05).

Table 25

Hierarchical Linear Models Using Stevens' OTL Measure to Predict 8th Grade Mathematics Achievement in TIMSS 2003 ($n_{unweighted} = 3780$)

	Model 1	Model 2	Model 3	Model 4	Model 5
Student-level variables					
Female Student		-3.35	-3.30	-3.06*	-3.16
		(2.66)	(2.66)	(2.66)	(2.54)
Student Age [b]		9.23*	9.23*	9.24***	8.62*
		(4.03)	(4.03)	(4.03)	(3.88)
Student SES		14.52***	14.52***	14.53***	14.71***
		(1.26)	(1.26)	(1.26)	(1.23)
Teacher-level variables					
Intercept	571.44***	561.14***	563.01***	547.42***	549.01***
	(3.05)	(19.65)	(19.59)	(19.12)	(18.40)
Class SES		63.90***	61.72***	60.58***	58.97***
		(8.07)	(8.17)	(7.76)	(7.38)
Female Teacher		3.78	3.44	5.04	6.19
		(4.98)	(4.96)	(4.75)	(4.42)
Teacher Age [a]					
Age 25-29		16.58	14.89	29.05	26.21
		(17.82)	(17.78)	(17.32)	(16.73)
Age 30-39		13.83	12.03	23.60	22.00
		(18.17)	(18.12)	(15.52)	(16.89)
Age 40-49		9.94	7.98	23.85	22.34
		(21.60)	(21.53)	(20.93)	(19.95)
Age 50-59		8.98	7.29	25.32	22.54
		(25.39)	(25.26)	(24.59)	(23.51)
Age 60 and over		9.43	3.57	25.45	20.03
		(39.84)	(39.87)	(38.15)	(36.60)
Teacher Experience [c]		-0.24	-0.22	-0.31	-0.25
		(0.70)	(0.69)	(0.66)	(0.62)
Number of Students in Class [c]		0.70	0.68	0.75	0.72
		(0.49)	(0.48)	(0.47)	(0.44)
Quality of Instructional Delivery OTL			0.42		0.31
			(0.37)		(0.32)
Instructional Time				7.06***	6.94***
				(1.95)	(1.89)
Random components					
Level-1 variance	5194.65	4994.30	4994.50	4993.72	5021.74
Level-2 variance	1098.07	400.62	393.13	344.81	317.65
Intraclass correlation coefficient (ICC)	0.17	0.07	0.07	0.06	0.06
% change in Level-2 variance compared to the null model	-	63.5%	64.2%	68.6%	71.1%

Notes: Standard errors in parentheses. Student weights were applied. [a] Reference group for Teacher Age—Under age 25. [b] Student Age was group mean centered. [c] Teacher Experience and Number of Students in Class were grand mean centered.
*p < 0.05, **p < 0.01, ***p < 0.001

Model 3 estimated the relationship between the *quality of instructional delivery* OTL measure and the students' mathematics performance on TIMSS 2007, adjusted for student-level and classroom-level covariates. Holding everything constant, there was no significant association between *quality of instructional delivery* OTL measure and the class-level mathematics performance ($b = 0.41$, $p > 0.05$). Using the *quality of instructional delivery* OTL measure as a predictor reduced the between-class variance by less than 1%, when compared to Model 2.

Model 4 estimated the relationship between *instructional time* and the mean mathematics performance of the post-*yutori* cohort, adjusted for student-level and classroom-level covariates. Controlling for everything else, there was no significant association between *instructional time* and the class-level mathematics performance ($b = -0.16$, $p > 0.05$). Using *instructional time* as a predictor did not further reduce the between-class variance, when compared to Model 2.

In Model 5, both the *instructional time* and *quality of instructional delivery* OTL measure were included as predictors of mean mathematics achievement on the TIMSS 2007. Controlling for *instructional time* and all other variables, there was no statistically significant correlation between the *quality of instructional delivery* OTL measure and the mean 8th graders' achievement on TIMSS 2007 ($p > 0.05$). Similarly, there was no statistically significant correlation between *instructional time* and the mean mathematics achievement on the TIMSS 2007 ($p > 0.05$) after holding the *quality of instructional delivery* OTL measure and all other variables constant. Including the *instructional time* and *quality of instructional delivery* OTL measure as predictors reduced the between-class variance by less than 1%, when compared to Model 2.

Table 26

Hierarchical Linear Models Using Stevens' OTL Measure to Predict 8th Grade Mathematics Achievement in TIMSS 2007 ($n_{unweighted}$ = 4784)

	Model 1	Model 2	Model 3	Model 4	Model 5
Student-level variables					
Female Student		-5.88	-5.88	-5.88	-5.55*
		(2.94)	(2.94)	(2.94)	(2.71)
Student Age [b]		10.57*	10.58*	10.57*	10.57*
		(4.25)	(4.25)	(4.25)	(4.25)
Student SES		20.40***	20.38***	20.39***	20.39***
		(1.51)	(1.52)	(1.51)	(1.51)
Teacher-level variables					
Intercept	567.66***	570.62***	571.61***	570.56***	576.67***
	(3.44)	(12.72)	(12.71)	(12.76)	(11.72)
Class SES		54.29***	54.88***	54.29***	53.57***
		(7.29)	(7.34)	(7.29)	(6.95)
Female Teacher		3.18	2.83	3.21	0.66
		(4.55)	(4.54)	(4.56)	(4.26)
Teacher Age [a]					
Age 25-29		-6.98	-7.97	-6.94	-13.16
		(10.85)	(10.83)	(10.87)	(9.94)
Age 30-39		4.40	2.74	4.42	-3.23
		(11.37)	(11.39)	(11.37)	(10.64)
Age 40-49		0.95	0.74	1.00	-2.06
		(15.18)	(15.09)	(15.19)	(14.29)
Age 50-59		7.94	8.40	8.07	5.60
		(19.14)	(19.06)	(19.25)	(17.77)
Age 60 and over		-0.23	-0.02	0.01	-3.79
		(22.81)	(22.68)	(23.08)	(21.87)
Teacher Experience [c]		-0.14	-0.18	-0.15	-0.29
		(0.57)	(0.57)	(0.57)	(0.54)
Number of Students in Class [c]		0.69	0.64	0.69	0.55
		(0.32)	(0.32)	(0.32)	(0.31)
Quality of Instructional Delivery OTL			0.41		0.58
			(0.37)		(0.35)
Instructional Time				-0.16	1.98
				(2.35)	(1.84)
Random components					
Level-1 variance	5839.94	5397.13	5396.90	5397.10	5411.33
Student SES slope		54.16	54.96	54.16	54.87
Level-2 variance	1381.24	534.50	527.96	534.64	515.95
Intraclass correlation coefficient (ICC)	0.19	0.09	0.09	0.09	0.09
% change in Level-2 variance compared to the null model	-	61.3%	61.8%	61.3%	62.6%

Notes: Standard errors in parentheses. Student weights were applied. [a] Reference group for Teacher Age—Under age 25. [b] Student Age was group mean centered. [c] Teacher Experience and Number of Students in Class were grand mean centered.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Interaction analysis of OTL-SES effects for pre-yutori cohort. Table 27 presents the OLS analyses results investigating the moderating effects of OTL measures on the Student SES-mathematics achievement relationship for the pre-*yutori* cohort (TIMSS 1999). For completeness and ease of comparison, I included the null model as Model A in Table 27. In Model B, I added all student-level and classroom-level covariates along with the *quality of instructional delivery* OTL measure and the student SES-*quality of instructional delivery* OTL interaction to the OLS model. Holding everything else constant, there was no evidence of the *quality of instructional delivery* OTL main effect and the student SES-*quality of instructional delivery* OTL interaction effect ($p > 0.05$).

In Model C, I examined the student SES-*instructional time* interaction effect on the mean mathematics performance of the pre-*yutori* cohort, adjusted for student-level and classroom-level covariates. Controlling for everything else, there was no significant student SES-*instructional time* interaction effect on students' mathematics performance ($b = 0.08$, $p > 0.05$).

In Model D, I included both student SES-*instructional time* and student SES-*quality of instructional delivery* OTL interaction effects on the mean mathematics performance of the pre-*yutori* cohort, adjusted for student-level and classroom-level covariates. Controlling for everything else, there was no significant student SES-*instructional time* interaction effect on students' mathematics performance ($b = 0.53$, $p > 0.05$). There was no evidence of a significant student SES-*quality of instructional delivery* OTL interaction effect on mean mathematics achievement ($b = 0.24$, $p > 0.05$).

Table 27

OLS Regression Models Investigating the Moderating Effects of OTL Measure on SES in Predicting 8th Grade Mathematics Achievement in TIMSS 1999 ($n_{unweighted}$ = 4606)

	Model A	Model B	Model C	Model D
Student-level variables				
Female Student		-6.61	-6.00*	-6.71
		(3.28)	(2.79)	(3.34)
Student Age		24.62***	23.86***	24.52
		(5.37)	(4.22)	(5.32)
Student SES		15.34***	15.49***	15.31***
		(1.94)	(1.57)	(1.94)
Student SES x Quality of Instructional Delivery OTL		0.24		0.24
		(0.23)		(0.22)
Student SES x Instructional Time OTL			0.08	0.53
			(1.10)	(1.21)
Teacher-level variables				
Intercept	579.60***	184.93*	191.50**	188.72
	(3.05)	(85.81)	(62.88)	(86.05)
Class SES		21.16	20.83*	21.29
		(13.31)	(9.99)	(13.53)
Female Teacher		-7.62	-4.32	-7.79
		(4.38)	(3.79)	(4.41)
Teacher Age [a]				
Age 25-29		-16.37	-8.93	-15.68
		(17.16)	(12.06)	(17.27)
Age 30-39		-11.18	-4.82	-10.22
		(19.23)	(14.47)	(19.58)
Age 40-49		-1.95	0.42	-0.86
		(22.05)	(17.67)	(22.91)
Age 50-59		-13.40	-7.14	-10.54
		(25.36)	(22.40)	(26.87)
Age 60 and over		-0.58	9.99	1.26
		(33.19)	(26.39)	(34.86)
Teacher Experience		0.06	0.22	0.01
		(0.91)	(0.73)	(0.95)
Number of Students in Class		1.49	1.33	1.41
		(1.01)	(0.77)	(1.06)
Quality of Instructional Delivery OTL		0.54		0.55
		(0.38)		(0.38)
Instructional Time			0.80	1.08
			(2.73)	(2.61)
Random components				
Level-1 variance	5587.49			
Level-2 variance	493.71			
Intraclass correlation coefficient (ICC)	0.08			
% change in Level-2 variance compared to the null model	-			

Notes: Standard errors in parentheses. Student weights were applied. Models B to D were estimated with using OLS regression with standard robust errors. [a] Reference group for Teacher Age—Under age 25.
*p < 0.05, **p < 0.01, ***p < 0.001

HLM analysis of cross-level OTL-SES effects for mid-yutori cohort. Table 28 presents the HLM analyses results investigating the moderating effects of OTL measures on the student SES-mathematics achievement relationship for the mid-*yutori* cohort (TIMSS 2003). Again, for completeness and ease of comparison, I included the null model as Model A in Table 28.

In Model B, I added all student-level and classroom-level covariates along with the *quality of instructional delivery* OTL measure and the cross-level student SES-*quality of instructional delivery* OTL interaction to the HLM model. Holding everything else constant, there was no evidence of the *quality of instructional delivery* OTL main effect and the student SES-*quality of instructional delivery* OTL interaction effect ($p > 0.05$). Model B further reduced between-class variance by 57%, when compared to the null model.

In Model C, I examined the cross-level student SES-*instructional time* interaction effect and the main *instructional time* effect on the mean mathematics performance of the mid-*yutori* cohort, adjusted for student-level and classroom-level covariates. Controlling for everything else, I observed a significant main effect of *instructional time* on class-level mathematics achievement (b = 7.25, SE = 1.96, $p < 0.001$), but I did not find a significant student SES-*instructional time* interaction effect on class-level mathematics performance ($b = -0.93$, $p > 0.05$). This finding suggested that the positive *instructional time* effect on mean achievement was the same for all mid-*yutori* cohort students. For a 1 standard deviation unit increase in their *instructional time* for mathematics class, students would be expected to improve their TIMSS mathematics assessment scores by 7 raw scale points. Model C further reduced between-class variance by 62%, when compared to the null model.

In Model D, I included both student SES-*instructional time* and student SES-*quality of instructional delivery* OTL cross-level interaction effects on the mean mathematics performance of the mid-*yutori* cohort, adjusted for student-level and classroom-level covariates. There was no evidence of a significant student SES-*quality of instructional delivery* OTL interaction effect ($b = 0.00$, $p > 0.05$) and no evidence of a significant *quality of instructional delivery* OTL main effect on mean mathematics achievement ($b = 0.32$, $p > 0.05$). Controlling for everything else, there was a significant *instructional time* main effect ($b = 7.31$, $p < 0.001$), but not a student SES-*instructional time* interaction effect on mean mathematics performance of mid-*yutori* cohort ($b = -1.00$, $p > 0.05$). Again, this suggested a positive association between *instructional time* and mean mathematics achievement for the mid-*yutori* cohort. Model D further explained 71% of between-class variance, when compared to the null model.

HLM analysis of cross-level OTL-SES effects for post-yutori cohort. Table 29 presents the HLM analyses results investigating the moderating effects of OTL measures on the student SES-mathematics achievement relationship for the post-*yutori* cohort (TIMSS 2007).

In Model B, I added all student-level and classroom-level covariates along with the *quality of instructional delivery* OTL measure and the cross-level student SES-*quality of instructional delivery* OTL interaction to the HLM model. Holding everything else constant, there was no evidence of the *quality of instructional delivery* OTL main effect nor the student SES-*quality of instructional delivery* OTL interaction effect ($p > 0.05$). Model B further reduced between-class variance by 62%, when compared to the null model.

Table 28

Hierarchical Linear Models Investigating the Moderating Effects of OTL Measure on SES in Predicting 8th Grade Mathematics Achievement in TIMSS 2003 ($n_{unweighted}$ *= 3780)*

	Model A	Model B	Model C	Model D
Student-level variables				
Female Student		-3.30	-3.10	-3.19
		(2.66)	(2.66)	(2.54)
Student Age [b]		9.23*	9.24*	8.63*
		(4.03)	(4.02)	(3.88)
Student SES		14.51***	14.74***	14.91***
		(1.26)	(1.28)	(1.25)
Student SES x Quality of Instructional Delivery OTL		0.04		0.00
		(0.21)		(0.21)
Student SES x Instructional Time OTL			-0.93	-1.00
			(0.93)	(0.95)
Teacher-level variables				
Intercept	571.44***	562.96***	546.84***	548.49***
	(3.05)	(19.59)	(19.14)	(18.41)
Class SES		61.66***	60.82***	59.14***
		(8.15)	(7.77)	(7.37)
Female Teacher		3.47	5.01	6.15
		(4.94)	(4.75)	(4.42)
Teacher Age [a]				
Age 25-29		14.83	29.74	26.86
		(17.74)	(17.36)	(16.75)
Age 30-39		12.04	24.24	22.60
		(18.10)	(17.54)	(16.90)
Age 40-49		8.01	24.47	22.92
		(21.51)	(20.95)	(19.96)
Age 50-59		7.35	25.84	23.00
		(25.26)	(24.61)	(23.51)
Age 60 and over		3.73	26.42	20.78
		(39.90)	(38.21)	(36.67)
Teacher Experience [c]		-0.22	-0.32	-0.25
		(0.69)	(0.66)	(0.62)
Number of Students in Class [c]		0.68	0.75	0.72
		(0.48)	(0.47)	(0.44)
Quality of Instructional Delivery OTL		0.42		0.32
		(0.36)		(0.32)
Instructional Time			7.25***	7.13***
			(1.96)	(1.91)
Random components				
Level-1 variance	5194.65	4994.50	4992.95	5019.80
Level-2 variance	1098.07	391.80	344.95	317.56
Intraclass correlation coefficient (ICC)	0.17	0.07	0.06	0.06
% change in Level-2 variance compared to the null model	-	57.2%	62.0%	71.1%

Notes: Standard errors in parentheses. Student weights were applied. [a] Reference group for Teacher Age—Under age 25. [b] Student Age was group mean centered. [c] Teacher Experience and Number of Students in Class were grand mean centered.
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

In Model C, I examined the cross-level student SES-*instructional time* interaction effect and the main *instructional time* effect on the mean mathematics performance of the post-*yutori* cohort, adjusted for student-level and classroom-level covariates. Controlling for everything else, I observed a significant cross-level, student SES-*instructional time* interaction effect (b = -2.80, SE = 1.38, $p < 0.05$), but I did not find a significant *instructional time* main effect on class-level mathematics achievement ($b = 1.25$, $p > 0.05$). Because the cross-level interaction effect is difficult to interpret, I graphed the interaction effect to better demonstrate it in context (see Figure 6). The student SES-*instructional time* interaction effect suggested that the effects of student SES on mathematic achievement depended on the amount of *instructional time* the student received in his/her mathematics class. As illustrated by the blue line in Figure 6, holding everything else constant, the achievement gap was the widest between low-SES students and high-SES students if both received 2 standard deviation units less than the mean *instructional time.* On the other hand, the achievement gap was the narrowest between low-SES students and high-SES students if both received 2 standard deviation units more than the mean *instructional time.* Put differently, this finding suggested that *yutori* reforms exacerbated the achievement gap in the post-*yutori* cohort. Model C further reduced between-class variance by 61%, when compared to the null model.

In Model D, I included both student SES-*instructional time* and student SES-*quality of instructional delivery* OTL cross-level interaction effects on the mean mathematics performance of the mid-*yutori* cohort, adjusted for student-level and classroom-level covariates. There was no evidence of a significant student SES-*quality of instructional delivery* OTL interaction effect ($b = 0.00$, $p > 0.05$) and no evidence of a significant *quality of instructional delivery* OTL main effect on mean mathematics achievement (*b* = 0.32, *p* > 0.05). Controlling for everything else, there was a significant cross-level effect student SES-*instructional time* interaction effect (b = -2.85, SE = 1.11, *p* < 0.05), but there was no significant *instructional time* main effect on class-level mathematics achievement (*b* = 2.97, *p* > 0.05). Similar to Model C, the student SES-*instructional time* interaction effect suggested that the effects of student SES on mathematic achievement depended on the amount of *instructional time* the student received in his/her mathematics class, even after controlling for *quality of instructional delivery* OTL measure and a host of student- and teacher-level covariates. Again, as illustrated by the blue line in Figure 7, holding everything else constant, the achievement gap was the widest between low-SES students and high-SES students if both received 2 standard deviation units less than the mean *instructional time.* This finding again suggested that *yutori* reforms exacerbated the achievement gap in the post-*yutori* cohort, even after controlling for *quality of instructional practices*. Model D further reduced between-class variance by 63%, when compared to the null model.

Sensitivity check of OTL-SES relationship at school level. Because of the student SES-*instructional time* interaction effect observed in HLM analyses, I performed a sensitivity check to explore the relationship between the two OTL measures and SES at school level. The purpose of this sensitivity check was to see if the changes in the *instructional time* and *quality of instructional delivery* OTL measures across cohorts were related to school socioeconomic segregation.

Table 29

Hierarchical Linear Models Investigating the Moderating Effects of OTL Measure on SES in Predicting 8th Grade Mathematics Achievement in TIMSS 2007 ($n_{unweighted}$ = 4784)

	Model A	Model B	Model C	Model D
Student-level variables				
Female Student		-5.84	-5.88	-5.52*
		(2.93)	(2.94)	(2.70)
Student Age [b]		10.61*	10.67*	10.68*
		(4.25)	(4.25)	(4.11)
Student SES		20.40***	20.26***	20.61***
		(1.51)	(1.49)	(1.39)
Student SES x Quality of Instructional Delivery OTL		0.13		0.04
		(0.23)		(0.22)
Student SES x Instructional Time OTL			-2.80*	-2.85*
			(1.38)	(1.11)
Teacher-level variables				
Intercept	567.66***	571.64***	569.76***	574.92***
	(3.44)	(12.70)	(12.80)	(11.83)
Class SES		54.78***	53.66***	52.67***
		(7.33)	(7.35)	(6.97)
Female Teacher		2.85	3.86	1.40
		(4.53)	(4.57)	(4.26)
Teacher Age [a]				
Age 25-29		-8.03	-7.10	-12.39
		(10.83)	(10.85)	(9.95)
Age 30-39		2.66	4.84	-2.00
		(11.39)	(11.37)	(10.68)
Age 40-49		0.70	1.35	-0.90
		(15.09)	(15.19)	(14.35)
Age 50-59		8.33	9.04	7.90
		(19.05)	(19.27)	(17.86)
Age 60 and over		0.33	4.64	2.11
		(22.61)	(23.28)	(22.08)
Teacher Experience [c]		-0.19	-0.17	-0.31
		(0.57)	(0.57)	(0.55)
Number of Students in Class [c]		0.64*	0.70	0.56
		(0.32)	(0.32)	(0.31)
Quality of Instructional Delivery OTL		0.37		0.53
		(0.38)		(0.36)
Instructional Time			1.25	2.97
			(2.47)	(1.89)
Random components				
Level-1 variance	5839.94	5396.90	5397.93	5410.34
Student SES slope		54.71	46.44	45.44
Level-2 variance	1381.24	526.72	534.82	516.13
Intraclass correlation coefficient (ICC)	0.19	0.09	0.09	0.09
% change in Level-2 variance compared to the null model	-	61.9%	61.3%	62.6%

Notes: Standard errors in parentheses. Student weights were applied. [a] Reference group for Teacher Age—Under age 25. [b] Student Age was group mean centered. [c] Teacher Experience and Number of Students in Class were grand mean centered.

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

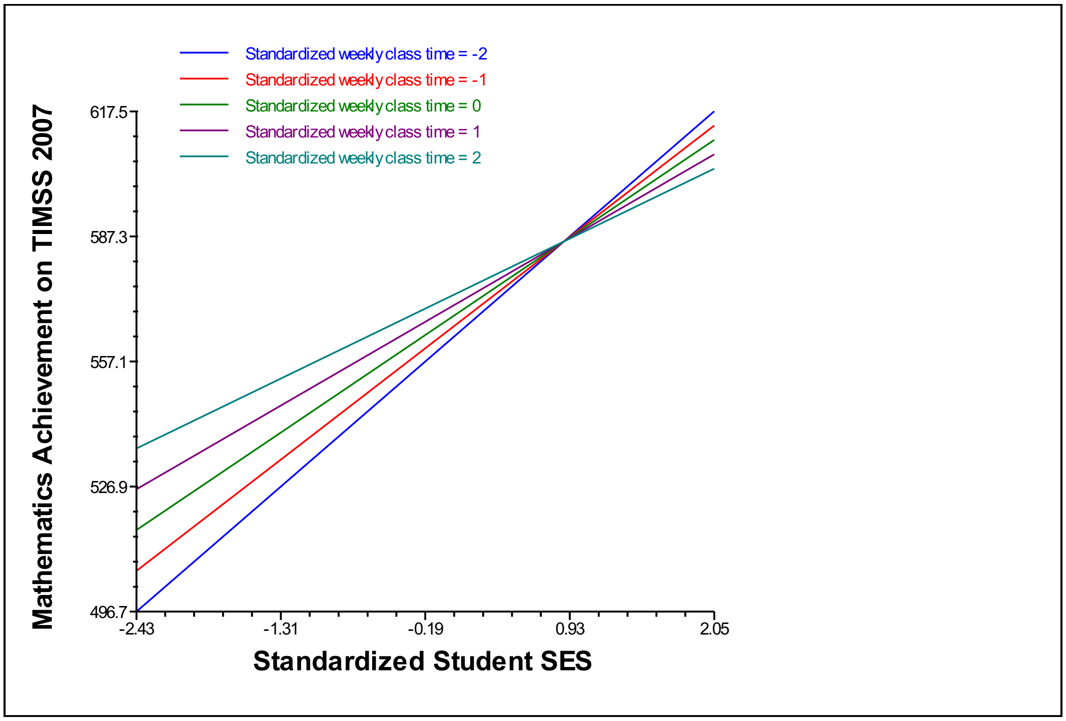


Figure 6. Student SES-instructional time interaction effect on student mathematics achievement in TIMSS 2007

The graph was generated from a HLM model with Student SES random slope while controlling for individual- and class-level variables.

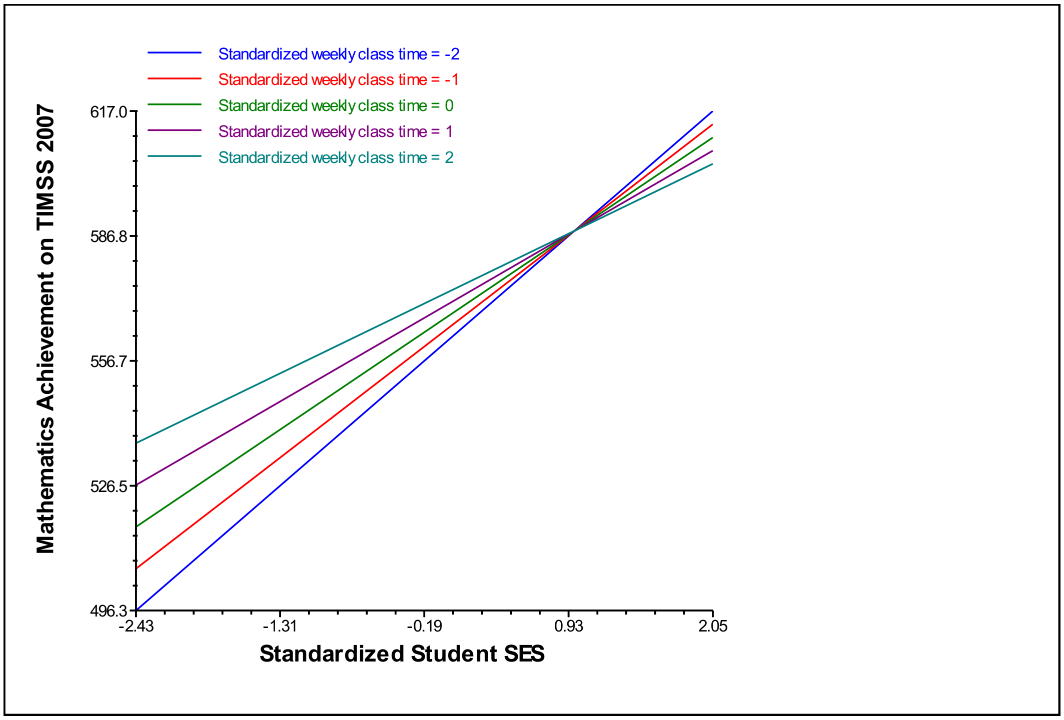


Figure 7. Student SES-instructional time interaction effect on student mathematics achievement in TIMSS 2007, controlling for quality of instructional delivery OTL

The graph was generated from a HLM model with Student SES random slope while controlling for quality of instructional delivery OTL and individual- and class-level variables.

I aggregated the student SES variable, *instructional time*, and *quality of instructional delivery* OTL measures at the school level and conducted the Pearson product-moment correlations with listwise deletion. No significant correlations were found between SES and *instructional time* at the school level (*r* ranged from 0.046 to 0.082). More variations were observed in the associations between the SES and *quality of instructional delivery* OTL: 0.125 for the pre-*yutori* cohort, 0.205 for the mid-*yutori* cohort, and -0.036 for the post-*yutori* cohort. Moreover, the association was statistically significant at 0.05 level for the mid-*yutori* cohort. This finding suggested that the *quality of instructional delivery* levels at a school seemed to be related to school-level socioeconomic segregation when *yutori* reforms were first implemented. On average, the mid-*yutori* cohort teachers working in socioeconomically advantaged schools reported using slightly more practices of consistent *quality of instructional delivery* OTL indicators when in teaching their mathematics classes.

Summary of Key Findings

To examine the relationship between the observed changes in OTL levels and students' mathematics achievement, I conducted classroom-level and multilevel analyses.

In the classroom-level analyses, I found that *yutori* reforms introduced between-cohort class-level achievement gaps by year via a substantial cut in instructional time. This between-cohort achievement gap would likely be diminished if all three cohorts were given an equal amount of weekly mathematics instructional time. I also found that the post-*yutori* teachers may have been able to narrow the achievement gap, which was imposed by the cut in instructional time, between them and the pre-*yutori* cohort by increasing their *quality of instructional delivery* OTL levels (see Table 21). Lastly, I found that *quality of instructional delivery* OTL played an important role in class-level mathematics achievement, even after accounting for *instructional time* and years of reform.

In the multilevel analyses, I found that the OTL measures—*quality of instructional delivery* and *instructional time*—were not significantly correlated with mean mathematics achievement for the pre-*yutori* cohort (i.e., those who participated in the TIMSS 1999) while controlling for student- and class-level variables.

For mid-*yutori* cohort (i.e., those who participated in the TIMSS 2003), I observed a significant association between *instructional time* and the class-level mathematics performance. Particularly, the HLM model predicts a 7-point gain on the TIMSS 2003 mathematics assessment for a student whose weekly mathematics class time increased by 1 standard deviation unit, controlling for everything else. This positive OTL-achievement relationship was the same for all students in the mid-*yutori* cohort.

For the post-*yutori* cohort (i.e., those who participated in the TIMSS 2007), I found a significant cross-level student SES-*instructional time* interaction effect, but I did not find any main OTL effects on mean mathematics achievement. The student SES-*instructional time* interaction effect suggested that the effects of student SES on mathematics achievement depended on the amount of *instructional time* the student received in his/her mathematics class. Holding everything else constant, the achievement gap was the widest between low-SES students and high-SES students if both received 2 standard deviation units less than the mean *instructional time*. Low-SES students benefitted from the added instructional time more than high-SES students. This could be because they did not have access to added educational resources that wealthier students did. Unfortunately, this finding suggested that *yutori* reforms exacerbated the achievement gap in the post-*yutori* cohort.

Lastly, the sensitivity check findings suggested that the *quality of instructional delivery* levels at a school seemed to be weakly correlated to school socioeconomic segregation when *yutori* reforms were first implemented.

In summary, I presented evidence of unintended class-level achievement gaps introduced by the *yutori* reforms among three *yutori* cohorts. I also presented some evidence in the post-*yutori* cohort of a significant decrease in the between-cohort achievement gap when teachers increased their *quality of instructional delivery* OTL levels. Lastly, I presented evidence of *yutori* reforms unintentionally exacerbating the achievement gap between low-SES and high-SES students in the post-*yutori* cohort via a cut in *instructional time* for mathematics class.

Chapter V

DISCUSSION

The introduction of the *yutori* reforms in 2002 was Japan's ambitious policy vision to reshape its schools. *Yutori* curricular reforms constituted a 30% cut in the curriculum of core academic subjects, the implementation of a 5-day school week, the introduction of IS classes, and the expansion of elective subjects (MEXT, 2002). The purpose of this study was to explore the effects of the *yutori* reforms on OTL, as defined by Stevens' (1993, 1996) multidimensional framework, and to examine how the changes in OTL may have subsequently affected Japanese 8th graders' mathematics achievement, as measured by the TIMSS 1999, 2003, and 2007. To address the research questions outlined in Chapter I, the study employed a mixed-methods multi-cohort study, combining analyses of archival documents and interview-based data with analyses of quantitative TIMSS data on OTL and student achievement in mathematics in selected years. Specifically, the study presented a multilevel analysis of curricular reforms over time in a single nation, using three TIMSS cohorts (pre-*yutori*, mid-*yutori*, post-*yutori*) of nationally representative 8th grade samples.

Limitations of Study

Although the findings of this study add to the literature of OTL and *yutori* reforms, the study was limited by the constraints of secondary data analysis—the work was limited to data already collected by the TIMSS and other researchers, and some of the information utilized may have been inadequate. For example, the TIMSS 2011 data could not be included in the study because there was a much small number of *content coverage* questions on the teacher survey.

The study focused on the mathematics domains of Japanese 8th graders. The results of the study should, therefore, not be generalized to other grade levels or subject areas. What the study was able to demonstrate was the interplay between the enacted *yutori* policies and OTL, with the latter having a consistent role in explaining 8th grade student achievement in mathematics. It should be noted that classroom-level achievement given by the TIMSS is more valid and reliable than individual student-level measures, due to the design employed during test administration and the estimation of the most plausible values for students.

While my interviews were a strength of the study that provided a picture of how *yutori* reforms were implemented in schools and classrooms, the accuracy of my reconstruction of this picture was dependent, to a great extent, on the accuracy of the memory of the interviewees. The retrospective interview and analysis of archival data were corroborated where possible to compensate for this limitation.

Lastly, as with any non-experimental study, the results presented in this study did not permit causal inferences from the estimated associations among the measured constructs in the quantitative results. By taking a longitudinal view, supported with a mixed-methods parallel analysis of data from multiple sources, the study attempted to obtain a comprehensive picture of how the reforms affected OTL and student achievement.

Key Findings

The three overarching findings of this study were: (a) *yutori* reforms as intended by MEXT were not implemented at schools and classes; (b) significant changes in classroom-level OTL measures, indicating reductions in *instructional time* but improvements in the *quality of instructional delivery*, were found to occur under the *yutori* reforms; and (c) *instructional time* was found to be positively associated with students' mathematics achievement under the *yutori* reforms, with the most socioeconomically disadvantaged students benefitting more in terms of achievement outcomes than those who were advantaged.

Yutori **Reforms Not Implemented as Intended**

Consistent with findings from prior research on *yutori* reforms (Bjork, 2015; Cave, 2016), the study presented evidence that the *yutori* reforms were not interpreted and implemented in schools and classrooms as MEXT had originally intended. The *yutori* reform effort fell well short of the transformation of educational practices in schools that the Education Ministry had expected. The "relaxed" and "liberal" school life emphasized in *yutori* was in direct conflict with the middle school teachers' deep-rooted beliefs in academic achievement and examination preparation. The literature on Japanese education showed that middle school instructors predominantly used an examination-focused, teacher-centered approach to teaching (Fukuzawa, 1994; LeTendre, 1994, 1995). Given that there was no significant change in the high school entrance examination system even after *yutori* reforms were officially implemented, LeTendre (2002) argued that this unchanged situation provided "impetus for teachers to ignore or diffuse the reforms" (p. 31). My interview findings and descriptive analysis results were in agreement with expectations set by prior research: both researchers whom I interviewed reported that middle school teachers continued to teach the pre-*yutori* curriculum; teachers reported, on TIMSS teacher surveys, that they covered topics that were not part of the nationally intended 8th grade mathematics curriculum. These examination-focused beliefs and practices were referred to as an institutional priority by the two researchers I interviewed. These institutionalized beliefs "tended to interlock and reinforce one another, making major innovation even more difficult" (Cave, 2016, p. 130).

The best example of the *yutori* innovations was the introduction of IS classes, which presented the biggest problem to the junior high school teachers. Because of their interdisciplinary nature, IS courses demanded much more knowledge and skills than most teachers possessed. Because of the great power entrusted to schools and teachers by MEXT to implement the reforms, teachers did not hesitate to change the schedule to fit their personal and institutional beliefs. According to the literature on curriculum reform implementation (Spillane, 2004; Tyack & Cuban, 1995), teachers tended to interpret the introduced innovations based on their existing beliefs or practices and tended to assimilate the new components into their present practices. For this reason, the existing teaching practices remained unchanged despite the introduction of a reform (Fullan, 2007; Sarason, 1971; Tyack & Cuban, 1995). These practices persist because “they make the complex practices of instruction predictable, controllable, and also labor-saving” (Tyack & Cuban, 1995, p. 86). Fullan (2007) suggested that “changes in belief and understanding are the foundation of achieving lasting reform” (p. 37).

In a review of research on standards-based reforms and accountability, Chatterji (2002) described the defining components of systemic reforms in the United States as:

1. the establishment of challenging standards in the academic disciplines that would define what students should know and be able to do;
2. alignment of curriculum and instruction, assessment and accountability, and teacher certification and professional development components, with new academic standards;
3. revamping school governance structures, allowing schools and teachers greater autonomy in how they organize instructional programs to achieve the high standards of student performance set by reforms at the local level.

The findings presented in the study suggested that the first two components of systemic reform initiatives were clearly missing in the way the *yutori* reforms were implemented. Without these two components in place, the *yutori* reforms were unlikely to promote comprehensive and coherent changes successfully in schools and classrooms.

Lastly, I found there was a changing leadership with possibly shifting priorities, both at the school level and the federal level, to lead the *yutori* reform effort. According to Researcher B, principals played an insignificant role in implementing the *yutori* reforms because they were usually in their post for 2 or 3 years before being transferred to another school. This rapid turnover contradicted the literature on educational change, which sees the principal as an important catalyst of change (Fullan, 2007; Hargreaves, Earl, Moore, & Manning, 2001). Likewise, there was instability in MEXT leadership. As a result, different policies were prioritized whenever there was a change in Education Ministry leadership positions. If those in authority positions do not receive adequate reform orientation, it is difficult for them to lead an effective system-wide reform (Fullan, 2007).

Changes in OTL Levels Under Yutori Reforms

Using content-validated OTL measures derived from TIMSS teacher surveys, I reviewed the changes in *content coverage, content emphasis, instructional time,* and *quality of instructional delivery* between cohorts (pre-*yutori*, mid-*yutori*, post-*yutori*). Based on the results of descriptive analyses, I did not find visible changes in *content coverage* and *content emphasis*. In other words, the mathematics topics that were covered and emphasized were similar between cohorts. However, I found significant differences in *instructional time* and *quality of instructional delivery* OTL between cohorts. Particularly, I found that instructional time for 8th grade mathematics classes decreased by at least 20% or by 40 minutes per week after the implementation of the *yutori* reforms. In other words, the *yutori* reforms contributed to a 20% drop in *instructional time* in 8th grade mathematics class. In contrast, I observed a significant increase of self-reported practices in *quality of instructional delivery* OTL after the *yutori* reforms were implemented. Particularly, the post-*yutori* cohort teachers were found to be the most prepared to teach mathematics and utilized the most variety in teaching strategies to meet their students' educational needs, followed by the mid-*yutori* cohort teachers, and lastly, the pre-*yutori* cohort teachers. The improvement in the standard score of the *quality of* *instructional delivery* OTL measure was more than four standard deviation units between the pre-*yutori* cohort teachers to the post-*yutori* cohort teachers.

These results seemed to contradict my interview finding that post-*yutori* teachers used free time to teach mathematics. On the contrary, the quantitative findings were actually corroborated by my interview analysis. The TIMSS question for the *instructional time* variable was: *How many minutes per week do you teach math to your math class?* Teachers were expected to provide the number of minutes they taught mathematics to their math class, according to the official schedule. This was corroborated by the teacher whom Researcher A interviewed: "We don't have enough time in the schedule to teach math, so we use the morning IS time for math." In other words, teachers were forced to use other free time to teach mathematics because there was not enough instructional time allocated for mathematics classes on the official schedule. Together, these two findings provided evidence that the *yutori* reforms decreased OTL as defined by *instructional time*.

The increase in *quality of instructional delivery* practices was cross-validated by a quote given by Researcher A: "…if they [teachers] noticed that the new curriculum didn't attach enough importance to a particular concept, a lot of times they would use old textbooks or supplementary activities to cover that material." According to Stevens' (1996) OTL framework, a notable *quality of instructional delivery* subindicator is: Teacher uses varied teaching strategies and practices to meet the educational needs of all students. Teachers appeared to be doing some compensatory actions to undo the effects of reduced *instructional time* under *yutori* reforms at their schools.

Association Between OTL and Student Achievement in Mathematics

My findings showed that *instructional time* and *quality of instructional delivery* OTL were significantly associated with mean mathematics achievement, even when one or the other OTL measure was held constant along with years of reform. In other words, *quality of instructional delivery* OTL was significantly correlated with mean mathematics performance at the classroom level after controlling for *instructional time* and *yutori* cohorts; but the reverse was also true. This speaks to the importance of both *instructional time* and *quality of instructional delivery* OTL in their contribution to academic achievement.

Moreover, the class-level gaps in mathematics achievement over time were found to be correlated with the reduction in *instructional time*, as outlined in the *yutori* reforms. There was no significant difference in mean mathematics performance between cohorts once the *instructional time* was controlled for. The finding of a negative association between the reduction in *instructional time* and mathematics achievement observed in the study was consistent with the extensive body of literature on dedicated instructional time and student achievement in academic subjects (Berliner, 1978, 1990; Denham & Lieberman, 1980; Fisher et al., 1981; Harnischfeger & Wiley, 1985; Karweit, 1985; Walberg & Frederick, 1982).

However, the between-classroom differences in mean mathematics performance became statistically significant between the pre-*yutori* cohort and the other two cohorts, when the *quality of instructional delivery* OTL was held constant. The pre-*yutori* cohort performed significantly better in mathematics assessments than the post-*yutori* cohort once *quality of instructional delivery* OTL was accounted for; this difference only emerged after the *quality of instructional delivery* OTL was held constant. This implicitly implies that the post-*yutori* cohort teachers appeared able to mitigate the adverse effect of reduced *instructional time* by adding more *quality of instructional delivery* practices to their teaching. This finding of a positive association between *quality of instructional delivery* and *academic achievement* was largely consistent with past research (Herman & Abedi, 2004; Saxe et al., 1999; Wang, 1998).

The HLM analysis results suggested that the OTL effect on mean mathematics achievement was overshadowed by more powerful student-level and classroom-level predictors, namely SES variables in all three cohorts. In fact, as documented in other studies, SES explained more than half of the variation in pupil achievement (Berliner, 2006; Blanden et al., 2005; Bradley & Corwyn, 2002; Davis-Kean, 2005). The findings related to SES were consistent with this line of research.

Moreover, the *instructional time* OTL measure was found to be significantly, positively correlated with the mean mathematics scores on the TIMSS 2003 even after controlling for SES and other covariates. This lent further evidence for the association between mean mathematics assessment scores on TIMSS 2003 and *instructional time*, considering it was only a year after the *yutori* reforms were introduced.

Lastly, I observed a significant cross-level, student SES-*instructional time* interaction effect in the post-*yutori* cohort, controlling for all other background variables of teachers and students. This cross-level interaction remained statistically significant even after controlling for *quality of instructional delivery* OTL. This cross-level interaction implied that the effects of student SES on mathematic achievement depended on the amount of *instructional time* the student received in his/her mathematics class. The observed achievement gap was the widest for the students in classrooms with the least amount of dedicated instructional time—low-SES students performed considerably lower than high-SES students under these conditions. Serendipitously, I believe Floraline Stevens (1993) offered the most meaningful interpretation of this interaction effect in one of her first writings about the multidimensional OTL framework:

> When middle-class students do not receive “good” teaching in the classroom, their education is supplemented in many cases by their parents’ ability to understand and teach them the skills and concepts at home, or they are taught by tutors. In contrast, poor and minority students in most instances are totally dependent upon what is offered by the teachers in their classrooms. Thus, these students’ ability to achieve academically at an accepted level is limited to what the teachers teach. (pp. 234-235)

This finding built on the work of Bjork (2015), Cave (2007), Park and Lee (2013), and Wada and Burnett (2011): the reduction in *instructional time*, outlined in the *yutori* reforms, exacerbated the achievement gaps between the economically advantaged and the economically disadvantaged students. The finding is particularly unsettling, given that students with advantaged family backgrounds also had more financial resources to invest in *juku* which, on average, cost more than $3,000 U.S. dollars per student annually (OECD, 2012) as a way to overcome the negative unintended effects imposed by the *yutori*.

Significance of the Study

The significance of the study is fourfold. Findings can help (a) inform the “standards crisis” debate around the *yutori* reforms using TIMSS data within Japan; (b) demonstrate a methodology for within-country examinations using ILSA data over time, to examine and interpret effects of natural experiments such as large-scale national reforms; (c) offer a direction to improve the design and validation of OTL measures from TIMSS teacher surveys; and (d) suggest ways to improve validity in interpreting ILSA results with reference to intra-nation, regional factors, moving away from misleading inter-country comparisons.

Inform the "Standards Crisis" Debate Around *Yutori* Reforms

From its beginning to its termination, the "standards crisis" debate around the *yutori* reforms persisted, despite a lack of reliable data for assessing students' scholastic achievement trends through domestic programs (Takayama, 2007, 2008; Tsuneyoshi, 2004). *Yutori* critics linked the decreasing hours of study, outlined by the reform, to a lowering of academic standards (Kariya & Shimizu, 2004; Tsuneyoshi, 2004), while others challenged the validity of such a claim (Takayama, 2008). However, this debate solely focused on the *attained curriculum* level and paid little or no attention to the *intended curriculum* and the *implemented curriculum*.

My findings confirmed the *yutori* critics' assertion that decreasing hours of study would lead to a lowering of academic achievement in students (Kariya & Shimizu, 2004; Tsuneyoshi, 2004). Differing from previous sensational discourses, I used multiple data sources to identify the features of the *yutori* reforms responsible for the unintended negative consequences on student achievement. I examined how *yutori* reforms were intended, enacted, and attained at three different levels. In particular, this study focused on OTL as the key construct because the *yutori* reforms explicitly highlighted a reduction in *instructional time*, which is one of the main components of OTL in Stevens' framework.

The study shed light on this debate by providing evidence that the between-cohort achievement gaps were associated with the reduction in *instructional time* at the class level. The study additionally highlighted the exacerbated achievement gaps between the economically advantaged and the economically disadvantaged students due to reduction in *instructional time*. These findings were likely the unintended consequences of the *yutori* reforms.

Demonstrate a Methodology for Within-Country Examination Using ILSA Data

This study offers a methodology for within-country examination using ILSA data over time, to examine and interpret effects of natural experiments such as large-scale, national reforms. First, the study demonstrated why a within-country examination provides a good basis for understanding relationships of contextual variables at different levels with student achievement outcomes in national education systems. ILSA programs are often used to make comparisons across systems in different nations. However, this between-country comparison relies heavily on many assumptions, namely that the outcome variables have the same meaning in every society and the tasks designed to measure the outcome variables are equally related to the experiences of students in every society (Mislevy, 1995). Further, there is the assumption that context variables (e.g., national policies, reforms) have the same effects on student achievement in the same way in all nations.

Another serious limitation of ILSA programs is their inability to communicate the extent of within-country variations in a single-number, a mean score (Mislevy, 1995). In the within-country examination presented here, I used multiple data sources to investigate the influence of *yutori* reforms at multiple levels in an effort to bypass this limitation and explain the variations in OTL and, subsequently, student performance in mathematics.

Besides the multilevel lens, the study cross-validated the quantitative results with the qualitative findings to address high-stakes, reform-related questions. The present study design speaks to how research programs on systemic reforms could be implemented using ILSA data to "support the growth of clearly defined systems in directions consistent with the mission of reforms, with timely, strategic information for all stakeholders" (Chatterji, 2002, p. 378). The information generated can then be used to drive reform policies that promote desired outcomes at all levels.

Offer a Direction to Improve the Design and Validation OTL Measures From TIMSS Teacher Surveys

The study offers a way to combine the existing TIMSS conceptual framework of educational opportunity with Stevens' multidimensional OTL framework (see Figure 3 in Chapter II). The current TIMSS tripartite framework mainly addresses the content coverage aspect of OTL. This study demonstrated ways to further unpack the OTL dimensions manifested in the implemented curriculum by drawing on Stevens' multidimensional framework in classrooms.

This study offers a way to merge two conceptual frameworks to improve the design and validation of OTL measures from TIMSS teacher surveys by using the Process Model (Chatterji, 2003). This iterative process yielded a theoretically meaningful OTL scale—*quality of instructional delivery*—allowing subsequent interpretations and uses in statistical modeling.

Suggest Ways to Improve Validity in Interpreting ILSA Results With Reference to Regional Factors

Because of the league table presentation format, ILSA reports tend to be interpreted as a "horse race" between countries (Pizmony-Levy, 2014). In typical practice today, those inter-country comparative score means are usually accepted at face value by the public and policymakers who, in turn, often generate incorrect score-based inferences or actions within given nations. Even Japan, a country known for excellent academic performance on ILSAs, was vulnerable to hasty interpretations and generalizations based on international rankings. Focusing solely on mean assessment scores or rankings can be subject to different degrees of invalidity when taken out of context (Chatterji, 2013), as this study has shown in the examination of Japanese 8th graders' performance on the TIMSS before and during the *yutori* reform periods.

Misinterpretations of ILSA reports can have negative consequences by spreading misinformation in larger national and societal contexts. When important contextual factors are ignored in generating the countries' average scores that are ranked, ILSA results will have less meaning or value in national contexts. Examining achievement performance without considering the learning opportunities provided to students misconstrues the meaningfulness of results.

To avoid unintended consequences of validity oversights, policymakers, the general public, and even the research community need to recognize the importance of interpreting ILSA results with reference to regional factors, such as characteristics of students, educational policies, and reforms that may be in effect at the time of testing (Backhoff, 2013; Engel & Feuer, 2014; Laurie, 2013; Wagemaker, 2013).

Directions for Future Research

There is little doubt that ILSA programs can provide important and powerful insights into limits and possibilities in national education contexts. This study suggested a way to supplement the TIMSS conceptual framework of educational assessment with Stevens' multidimensional OTL framework. The validated *quality of instructional delivery* OTL variable has provided invaluable information about how the *yutori* reforms were implemented in classrooms. I hope the IEA will consider incorporating the multidimensional OTL constructs in future TIMSS survey items and assessments as this could improve future examinations of within-nation policy issues.

As alluded to earlier in this paper, the *yutori* reforms did not possess the essential components of systemic standard-based reforms, as observed by researchers in U.S. education contexts. This study revealed a mismatch between the *intended yutori curriculum* and the *implemented yutori curriculum*. The following recommendations that stem from this study could shed further light on future educational reform efforts in Japan:

1. Future reform efforts should be conceptualized based on the essential components of systemic reforms and should be accompanied by a systemic research program designed to track large-scale change over time.
2. Teachers, as well as other stakeholders, should have a voice in future reform efforts to promote buy-in and commitment from these stakeholders.
3. Strong leadership at the national level, the system level, and the school level could lead to more effective system-wide reform implementation.

REFERENCES

Abedi, J., Courtney, M., Leon, S., Kao, J., & Azzam, T. (2006). *English language learners and math achievement: A study of opportunity to learn and language accommodation* (Technical Report 702). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Adelman, E., Moore, A.-M., & Manji, S. (2011). *Using opportunity to learn and early grade reading fluency to measure school effectiveness in Mozambique*. Retrieved from http://www.epdc.org/sites/default/files/documents/OTL Mozambique.pdf

Alkin, M., Doby, W., & Lindheim, E. (1990). Ten schools program evaluation reports case studies: 1988-1989 update. Los Angeles, CA: Los Angeles Unified School District (LAUSD).

Asahi. (2004a, December 7). Gakuryoku toppu katsuraku no shogeki [Shocking fall from the academic top]. *Asahi Shimbun*, p. 2.

Asahi. (2004b, December 7). Sugaku ronjutsu yowai nihon [Japan weak in maths and essay-type questions]. *Asahi Shimbun*, p. 5.

Backhoff, E. (2013). Validity issues in international large scale assessment (ILSA) programs: Thoughts for developing countries. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 233-150). Bingley, UK: Emerald Publishing Limited.

Baker, D. P., Fabrega, R., Galindo, C., & Mishook, J. (2004). Instructional time and national achievement: Cross-national evidence. *Prospects, 34*(3), 311-334.

Berliner, D. C. (1978). Allocated time, engaged time and academic learning time in elementary mathematics instruction. Retrieved from https://eric.ed.gov/?id=ED171539

Berliner, D. C. (1981). Academic learning time and reading achievement. *Comprehension and Teaching: Research Reviews*, 203-226.

Berliner, D. C. (1990). What's all the fuss about instructional time. In M. Ben-Peretz & R. Bromme (Eds.), T*he nature of time in schools: Theoretical concepts, practitioner perceptions* (pp. 3-35). New York, NY: Teachers College Press.

Berliner, D. C. (2006). *Our impoverished view of educational research*. New York, NY: Teachers College Press.

Bjork, C. (2009). Local implementation of Japan's integrated studies curriculum: A preliminary analysis of efforts to decentralize the curriculum. *Comparative Education*, *45*(1), 23–44.

Bjork, C. (2015). *High-stakes schooling: What we can learn from Japan's experiences with testing, accountability, and education reform*. Chicago, IL: University of Chicago Press.

Bjork, C., & Tsuneyoshi, R. (2005). Education reform in Japan: Competing visions for the future. *Phi Delta Kappan, 86*(8), 619-626.

Blanden, J., Gregg, P., & Machin, S. (2005). Social mobility in Britain : Low and falling. *CentrePiece*, *Spring*, 18-20.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1-12.

Bloom, B. S. (1974). Time and learning. *American Psychologist, 29*(9), 682-688.

Bogdan, R. C., & Biklen, S. K. (1998). *Qualitative research for education: An introduction to theory and methods* (3rd ed.). Boston, MA: Allyn and Bacon.

Borg, W. R. (1980). Time and school learning. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 33-72). Washington, DC: National Institute of Education.

Boscardin, C. K., Aguirre-Muñoz, Z., Chinen, M., Leon, S., & Shin, H. S. (2004). Consequences and validity of performance assessment for English learners: Assessing opportunity to learn (OTL) in Grade 6 language arts (CSE Report 635). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualtative Research Journal, 9*(2), 27-40.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*(1), 371-399.

Braun, H. (2014). Merits of international assessmens. *Quality Assurance in Education, 22*(4), 332-334.

Brewer, D. J., & Stasz, C. (1996). *Enhancing opportunity to learn measures in NCES data*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/reprints/RP581.html

Brophy, J., & Good, T. (1986). Teacher-effects results. In M. Wittrock (Ed.), *Handbook of Research on Teaching* (pp. 328-375). New York, NY: Macmillan.

Carnoy, M., Khavenson, T., Loyalka, P., Schmidt, W. H., & Zakharov, A. (2016). Revisiting the relationship between international assessment outcomes and educational production: Evidence from a longitudinal PISA-TIMSS Sample. *American Educational Research Journal, 53*(4), 1054-1085.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*(8), 723-733.

Carroll, J. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher, 18*(1), 26-31.

Cave, P. (2003). Japanese educational reform: Developments and prospects at primary and secondary level. In R. Goodman & D. Phillips (Eds.), *Can the Japanese change their education system?* Oxford, UK: Symposium Books.

Cave, P. (2007). *Primary school in Japan: Self, individuality, and learning in elementary education*. London, UK: Routledge.

Cave, P. (2016). *Schooling selves: Autonomy, interdependence, and reform in Japanese junior high education*. Chicago, IL: University of Chicago Press.

Central Council on Education. (1972). *Basic guidelines for the reform of education of development of an integrated education system suited for contemporary society*. Tokyo, Japan: Author.

Central Council on Education. (1997). *Second report on the model for Japanese education in the perspective of the 21st century*. Tokyo, Japan: Author.

Cetola, C., DeStefano, J., Schuh Moore, A., & Adelman, E. (2010). *School effectiveness: Improving the use of financial investments in education*. Retrieved from http://www.epdc.org/sites/default/files/documents/School Effectiveness Improving the Use of Financial Investments in Education.pdf

Chatterji, M. (in press). Designing assessment for multidisciplinary constructs and apllications: A user-centered methodology. New York, NY: Guilford.

Chatterji, M. (2002). Models and methods for examining standards-based reforms and accountability initiatives: Have the tools of inquiry answered pressing questions on improving schools? *Review of Educational Research, 72*(3), 345-386.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston, MA: Allyn and Bacon.

Chatterji, M. (2013). Insights, emerging taxonomies, and theories of action towards improving validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 273-308). Bingley, UK: Emerald Group Publishing.

Chatterji, M., & Lin, M. (2018). Designing non-cognitive construct measures that improve mathematics achievement in grade 5-6 learners: A user-centered approach. *Quality Assurance in Education: An International Perspective, 26*(1), 70-100.

Cogan, L. S., & Schmidt, W. H. (2015). The concept of opportunity to learn (OTL) in international comparisons of education. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 207–216). Basel, Switzerland: Springer International.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Creswell, J. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage.

Creswell, J., & Plano-Clark, V. (2018). *Designing and conducting mixed-methods research*. *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology, 19*(2), 294-304.

Denham, C., & Lieberman, A. (1980). *Time to learn*. Washington, DC: National Institute of Education.

Desimone, L., Smith, T. M., & Phillips, K. (2013). Linking student achievement growth to professional development participation and changes in instruction: A longitudinal study of elementary students and teachers in Title I schools. *Teachers College Record, 115*(5), 1-46.

DeStefano, J., Adelman, E., & Schuh Moore, A. (2010). *Using opportunity to learn and early grade reading fluency to measure school effectiveness in Nepal executive summary*. Retrieved from http://www.equip123.net/docs/e2-School_ Effectiveness_Nepal-CS.pdf

DeStefano, J., & Elaheebocus, N. (2008). Using opportunity to learn and early grade reading fluency to measure school effectiveness in Woliso, Ethiopia. Retrieved from http://pdf.usaid.gov/pdf_docs/pnadz787.pdf

Elliott, S. N., & Bartlett, B. J. (2016). Opportunity to Learn. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935291.001.0001/oxfordhb-9780199935291-e-70.

Engel, L., & Feuer, M. J. (2014). Five myths about international large scale assessments (ILSAs). *Quality Assurance in Education, 22*(4), 326-328.

Feuer, M. J. (2013). Validity issues in international large-scale assessment programs: "Truth and consequences." In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 197-216). Bingley, UK: Emerald Publishing.

Fisher, C. W., Berliner, D. C., Filby, N. N., Marliave, R., Cahen, L. S., & Dishaw, M. M. (1981). Teaching behaviors, academic learning time, and student achievement: An overview. *The Journal of Classroom Interaction, 17*(1), 2-15.

Floden, R. E. (2002). The measurement of opportunity to learn. In *Methodological advances in cross-national surveys of educational achievement* (pp. 243-278). Washington, DC: National Academy Press.

Floden, R. E., Porter, A., Schmidt, W. H., Freeman, D. J., & Schwille, J. R. (1981). Responses to curriculum pressures: A policy-capturing study of teacher decisions about content. *Journal of Educational Psychology, 73*(2), 129-141.

Foy, P., Arora, A., & Stanco, G. (2013). *TIMSS 2011 user guide for the international database*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Foy, P., & Joncas, M. (2000). TIMSS sample design. In M. O. Martin, K. D. Gregory, S. E. Stemler (Ed.), *TIMSS 1999 Technical Report* (pp.29-48). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Foy, P., & Joncas, M. (2004). TIMSS 2003 sampling design. In M. O. Martin, I. V. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 Technical Report* (pp. 109-124). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Foy, P., & Olson, J. (2009). *TIMSS 2007 international database and user guide*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Fukuzawa, R. (1994). The path to adulthood according to Japanese middle schools. *The Journal of Japanese Studies, 20*, 61-86.

Fullan, M. (2007). *The new meaning of educational change*. New York, NY: Teachers College Press.

Gamoran, A. (1987). Instruction and the effects of schooling. Presented at the annual meeting of the American Sociological Association, Chicago, IL.

Gamoran, A., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*(4), 325-338.

Gillies, J., & Quijada, J. J. (2012). *Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries.* Washington, DC: United States Agency International Development.

Goldenberg, C., & Gallimore, R. (1991). Local knowledge, research knowledge, and educational change: A case study of early Spanish reading improvement. *Educational Researcher, 20*(8), 2-14.

Gonzalez, E., & Miles, J. (2001). *TIMSS 1999: User guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Guest, G. S., Namey, E. E., & Mitchell, M. L. (2013). *Collecting qualitative data: A field manual for applied research*. Thousand Oaks, CA: Sage.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*(3), 323-336.

Hargreaves, A., Earl, L., Moore, S., & Manning, S. (2001). *Learning to change: Learning beyond subjects and standards*. San Francisco, CA: Jossey-Bass.

Harnischfeger, A., & Wiley, D. E. (1985). Origins of active learning time. In C. W. Fisher & D. C. Berliner (Eds), *Perspectives on instructional time* (pp. 133-156). New York, NY: Longman.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. Abingdon, UK: Routledge.

Herman, J. L., & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics*. (CRESST Tech. Report 633). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., Klein, D. C., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 19*(4), 16-24.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Hox, J. J. (2010). *Multilevel analysis*. New York, NY: Routledge.

Husen, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vols. 1 & 2). New York, NY: John Wiley.

Japan Committee for Economic Development. (1984). A proposition from businessman for education reform: In pursuit of creativity, diversity, and internationality. In J. M. V. Edward & R. Beauchamp Jr. (Eds.), *Japanese education Since 1945* (pp. 284-291). New York, NY: Routledge.

Japan Society of Mathematical Education. (2000). *Mathematics program in Japan: Elementary, lower secondary and upper secondary schools.* Tokyo, Japan: Author.

Joncas, M. (2008). TIMSS 2007 sample design. In J. Olson, M. O. Martin, & I. V. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 77-92). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Kariya, T. (2009). From credential society to "learning capital" society: A rearticulation of class formation in Japanese education and society. In H. Ishida & D. H. Slater (Eds.), *Contemporary Japan structures sorting and strategies* (pp. 87-113). New York, NY: Routledge.

Kariya, T. (2010). *Education reform and social class in Japan.* London, UK: Routledge.

Kariya, T., & Rappleye, J. (2010). The twisted, unintended impacts of globalization on Japanese education. In E. Hannum, H. Park, & Y. G. Butler (Eds.), *Globalization demographic change and educational challenges in East Asia* (pp. 17-63). Bingley, UK: Emerald Publishing.

Kariya, T., & Shimizu, K. (2004). *Sociology of academic ability*. Tokyo, Japan: Iwanami Shoten.

Karweit, N. (1985). Should we lengthen the school term? *Educational Researcher, 14*(6), 9-15.

Kifer, E., & Burstein, L. (1992). Concluding thoughts: what we know, what it means. In L. Burstein (Ed.), *The IEA study of mathematics III: student growth and classroom processes* (pp. 329-342). Oxford, UK: Pergamon Press.

Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In *Handbook of accessible achievement tests for all students* (pp. 99-129). New York, NY: Springer

Kurz, A., Elliott, S. N., Kettler, R. J., & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment, 19*(3), 159-184.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *The Journal of Special Education*, *44*(3), 131-145.

Labuschagne, A. (2003). The qualitative report on qualitative research: Airy fairy or fundamental? *The Qualitative Report, 8*(1), 100-103.

Lafontaine, D., Baye, A., Vieluf, S., & Monseur, C. (2015). Equity in opportunity-to-learn and achievement in reading: A secondary analysis of PISA 2009 data. *Studies in Educational Evaluation, 47*, 1-11.

Laurie, R. (2013). Applying Feuer's validation framework in a Canadian context: A look at international large scale assessment programs. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 263-272). Bingley, UK: Emerald Publishing.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.

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

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What is tested, what's taught? *Journal of Educational Measurement, 18*(2), 85-96.

LeTendre, G. (1994). Guiding them on: Teaching, hierarchy, and social organization in Japanese middle schools. *The Journal of Japanese Studies, 20*(1), 37-59.

LeTendre, G. (1995). Disruption and reconnection: Counseling young adolescents in Japanese schools. *Educational Policy, 9*, 169-184.

LeTendre, G. (2002). Setting national standards: Educational reform, social change, and political conflict. In G. DeCoker (Ed.), *National standards and school reform in Japan and the United States* (pp. 19-32). New York, NY: Teachers College Press.

Martin, M. O. (Ed.). (2005). *TIMSS 2003 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision & Curriculum Development (ASCD).

McDonnell, L. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305-322.

McDonnell, L. M., Burstein, L., Ormseth, L., Catterall, J. S., & Moody, D. M. (1990). *Discovering what schools really teach: Designing improved coursework indicators.* Santa Monica, CA: Rand Corporation.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185-196.

Merriam, S. B. (2009). *Qualitative research : a guide to design and implementation*. San Francisco, CA: Jossey-Bass.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2000). *Toward a culturally oriented nation: Japnese government policies in education, science, sports, and culture*. Tokyo, Japan: Author.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2001). *Educational reform for the 21st century.* White paper. Tokyo, Japan: MEXT.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2002). *Creating the dreams for Japan's tomorrow: The role of the Ministry of Education, Culture, Sports, Science, and Technology*. Tokyo, Japan: Author.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2003). *School in the new era: Elementary and secondary education reform in progress*. Tokyo, Japan: Author..

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2008). 2008 students and extracurricular activities survey. Retrieved from http://www.mext.go.jp/b_menu/houdou/20/08/08080710.htm

Mimiduka, H. (2007). Determinants of children's academic achievement in primary education. *The Journal of Educational Sociology, 80*, 3-39.

Mislevy, R. (1991). Randomization based inference about examinees in the estimation of item parameters. *Psychometrika, 56*(2), 177-193.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*(4), 419-437.

Mo, Y. (2008). *Opportunity to learn, engagement, and science achievement: Evidence from TIMSS 2003 data.* (Doctoral dissertation, Virginia Tech).

Mullis, I. V., & Martin, M. O. (2013). *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., Gonzalez, E., & Chrostowski, S. (2004). *TIMSS 2003 International Mathematics Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., Gonzalez, E., Gregory, K., & Garden, R. (2000). *TIMSS 1999 international mathematics report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth grade*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nikkei. (2004, December 5). Nihon gakuryoku ohabani teika [Japanese academic achievement in big decline]. *Nikkei Shimbun*, p. 42.

Novak, J. D. (2008). T*he Theory Underlying Concept Maps and How to Construct and Use Them (Technical Report IHMC CmapTools 2006-01*). Retrieved from https://web.stanford.edu/dept/SUSE/projects/ireport/articles/concept_maps/The%20Theory%20Underlying%20Concept%20Maps.pdf

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2009). *Education at a glance 2009*. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2011). OECD economic surveys: Japan 2011. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2012). *Strong performers and successful reformers in education lessons from PISA for Japan*. Paris, France: Author.

Park, H., & Lee, Y.-J. (2013). Growing educational inequality in Japan during the 2000s. In G. DeCoker & C. Bjork (Eds.), *Japanese education in an era of globalization* (pp. 131-146). New York, NY: Teachers College Press.

Patton, M. Q. (2002). *Qualitative evaluation and research methods*. Thousand Oaks, CA: Sage.

Pizmony-Levy, O. (2014). Back to the future on international assessments. *Quality Assurance in Education, 22*(4), 321-322.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health, 30*, 459-467.

Porter, A. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis, 13*(1), 13-29.

Porter, A. (1993). Research news and comment: School delivery standards. *Educational Researcher, 22*(5), 24-30.

Porter, A. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A., & Smithson, J. (2001). *Defining, developing, and using Curriculum indicators*. Retrieved from Consortium for Policy Research in Education Website: http://www.cpre.org/sites/default/files/researchreport/788_rr48.pdf.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, T., & Congdon, R. (2000). *HLM 5*. Chicago, IL: Scientific Software International.

Rohlen, T. P. (1983). *Japan's high schools*. Berkeley, CA: University of California Press.

Rohlen, T. P. (1988). Education in Japanese society. In T. P. Rohlen & D. I. Okimoto (Eds.), *Inside the Japanese system: Readings on contemporary society and political economy*. Stanford, CA: Stanford University Press.

Sarason, S. (1971). *The culture of the school and the problem of change*. Boston, MA: Allyn and Bacon.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction, 17*(1), 1-24.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness* (1st ed.). Bingley, UK: Emerald Publishing.

Scheerens, J., Luyten, H., Steen, R., & Luyten-De Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness*. Enschede, The Netherlands: University of Twente, Department Educational Organisation and Management.

Schmidt, W. H., & Burroughs, N. A. (2013). *PISA 2012 assessment and analytical framework mathematics, reading, science, problem solving and financial literacy* (Vol. 39). Paris, France: OECD Publishing.

Schmidt, W. H., & Burstein, L. (1993). Concomitants of growth in mathematics achievement during the Population A school year. In L. Burstein (Ed.), *The IEA study of mathematics III: student growth and classroom processes* (pp. 309-326). Oxford, UK: Pergamon.

Schmidt, W. H., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook on education policy research* (pp. 541-559). New York, NY: Routledge.

Schmidt, W. H., & McKnight, C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis, 17*(3), 337-353.

Schmidt, W. H., McKnight, C., Cogan, L., & Jakwerth, P. (1999). Facing the consequences. In W. H. Schmidt & C. McKnight (Eds.), *Using TIMSS for a Closer Look at US Mathematics and Science Education*. Washington, DC: Kluwer Academic.

Schmidt, W. H., McKnight, C., Houang, R., Wang, H., Wiley, D. E., Cogan, L. S., & Wolfe, R. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.

Schmidt, W. H., Wolfe, R., & Kifer, E. (1992). The identification and description of student growth in mathematics achievement. In L. Burstein (Ed.), *The IEA study of mathematics III: Student growth and classroom processes* (pp. 59-99). Oxford, UK: Pergamon.

Schmidt, W. H., Zoido, P., & Cogan, L. (2013). *Schooling matters: Opportunity to learn in PISA 2012*. Paris, France: OECD Publishing.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51*(4), 455-498.

Smithson, J. L., & Collares, A. C. (2007). Alignment as a predictor of student achievement gains. Presented at the Annual meeting of the American Educational Research Association, Chicago, IL.

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

Spillane, J. (2004). *Standards deviation: How schools misunderstand education policy*. Cambridge, MA: Harvard University Press.

Stedman, L. C. (1994). Incomplete explanations: The case of U.S. performance in the international assessments of education. *Educational Researcher, 23*(7), 24-32.

Stevens, F. (1993). Applying an opportunity-to-learn conceptual framework to the investigation of the effects of teaching practices via secondary analyses of multiple-case-study summary data. *Journal of Negro Education*, *62*, 232-248.

Stevens, F. (1996). *Opportunity to learn science: Connecting research knowledge to classroom practices*. Washington, DC: Office of Educational Research and Improvement.

Stevenson, H., & Stigler, J. (1992). *The learning gap*. New York, NY: Summit Books.

Strauss, A., & Corbin, J. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Suter, L. E. (2017). How international studies contributed to educational theory and methods through measurement of opportunity to learn mathematics. *Research in Comparative and International Education, 12*(2), 174-197.

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Takayama, K. (2007). A nation at risk crosses the Pacific: Transnational borrowing of the U.S. crisis discourse in the debate on education reform in Japan. *Comparative Education Review, 51*(4), 423-446.

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*(4), 387-407.

Travers, K., Garden, R., & Rosier, M. (1989). Introduction to the study. In D. Robitaille & R. A. Garden (Eds.), *The IEA Study of mathematics II: Contexts and outcomes of school mathematics*. Oxford, UK: Pergamon.

Travers, K., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula.* Oxford, UK: Pergamon.

Tsuneyoshi, R. (2004). The new Japanese educational reforms and the achievement "crisis" debate. *Educational Policy, 18*(2), 364-394.

Tyack, D., & Cuban, L. (1995). Tinkering toward Utopia. *Social Service Review, 71*(3), 503-506.

von Davier, M., Gonzalez, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9-36.

Wada, M., & Burnett, B. M. (2011). Yutori Kyoiku and the uncertainty of recent neo-liberal reform in Japanese higher education. *Bulletin of Center for the Research and Support of Educational Practice*, 69-78.

Wagemaker, H. (2013). International large scale assessment (ILSA) programs and the challenges of consequential validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 217–232). Bingley, UK: Emerald Publishing.

Walberg, H. J. (1980). *A psychological theory of educational productivity.* Washington, DC: National Institute of Education.

Walberg, H. J. (1986). Synthesis of research on time and learning. *Education Leadership, 45*(6), 76-85.

Walberg, H. J., & Frederick, W. C. (1982). Instructional time and learning. In *Encyclopedia of educational research* (Vol. 2, pp. 917-924). New York, NY: Free Press.

Walker, D. F., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research, 44*(1), 83-111.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*(3), 137-156.

Wiley, D. E. (1990). *Opportunity to learn: A briefing for the Advisory Council on Education Statistics, National Center for Education Statistics*. Washington, DC: National Center for Education Statistics.

Wiley, D. E., & Harnischfeger, A. (1974). Explosion of a myth: Quantity of schooling and exposure to instruction, major educational vehicles. *Educational Researcher, 3*(4), 7-12.

Wilkinson, R., & Pickett, K. (2011). *The spirit level: Why greater equality makes societies stronger*. New York, NY: Bloomsbury.

Winfield, L. F. (1987). Teachers' estimates of test content covered in class and first-grade students' reading achievement. *The Elementary School Journal, 87*(4), 437-454.

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education, 62*(3), 288.

Yin, R. K. (2014). *Case study research: Design and methods. Essential guide to qualitative methods in organizational research* (5th ed.). Thousand Oaks, CA: Sage.

Yomiuri. (2004a, December 7). Kodomono gakuryoku kishingo [Yellow signal to children's academic achievement]. *Yomiuri Shimbun*.

Yomiuri. (2004b, December 7). Nihon no 15 sai gakuryoku toppu kanraku [Japan's 15-year-olds drop from the academic best]. *Yomiuri Shimbun*, p. 1.

Yomiuri. (2004c, December 7). Seiseki joi ko wa donna kyoiku [What kind of education at the top performing schools?]. *Yomiuri Shimbun*, p. 18.

Yoon, B., Burstein, L., Gold, K., Chen, Z., & Kim, K. (1990). Validating teachers' reports of content coverage: An example from secondary school mathematics. Presented at the Annual meeting of the National Council of Measurement in Education, Boston, MA.

Appendix A

Interview Guide (Part I and Part II)

Thank you for agreeing to participate in this study. As you know, I am Meiko Lin, a doctoral candidate at Teachers College, Columbia University in New York. I am the principal investigator of this study.

This study is being done to explore the effects of *yutori* reforms on opportunity-to-learn. This study also examines how this reform implementation, through a host of factors, affected students' opportunity to learn mathematics and subsequently students' achievement in mathematics.

Our interview today will last approximately an hour to two hours during which I will be seeking your personal impressions and knowledge about the *yutori* reforms based on your own experience. As a research participant, I would like to discuss your rights and receive your official permission for participation.
(*Present the Informed Consent Form, along with participant rights sheet and permission for audio-taping*)

Thank you very much for giving me the permission to conduct this interview. If there are any question you do not wish to answer, please let me know and we will skip to the next question. I want to clarify that your participation in this study is completely free and voluntary; you may refuse to respond to any questions; and you may discontinue with the study at any time. Do you have any questions about this research study and your role as a participant before we begin?

Before we begin the interview, I'd like to first operationally define the *yutori* reforms for this study so that we can be sure we are referring to the same thing. In this study, the *yutori* reforms refer to a Japanese education policy, which was implemented in 2002, aimed to reduce academic pressure and increase students' motivation to learn, creativity, and critical thinking. The *yutori* reforms consisted of six components: (1) Shortened school week, (2) Modification to the Course of Study, (3) Introduction of the Integrated Studies course, (4) Expansion of the elective courses, (5) Innovative pedagogy, and (6) Supportive teacher guidance/class management. Do you have any questions about this definition?
(*Provide detailed descriptions of the six components of the yutori reforms if needed.*)

Interview Guide (Part II)

Interviewee Name: ______________________________

Interviewer Name: ______________________________

Date of Interview: ______________________________

Time of Interview: ______________________________

Location of Interview: ______________________________

Interview Format: ______________________________

Interview Notes

<u>Interview Questions (for Researchers)</u>

According to your publications, you observed several schools after the yutori reforms were implemented.

1. How were you introduced to the *yutori* reforms?
2. Could you take me step-by-step through how the *yutori* reforms were implemented at the schools you observed?
3. What changes did you see in classrooms and schools after the *yutori* reforms were implemented?
4. Can you give an example of how *yutori* reforms shaped curriculum and classroom instructions?
5. How did students and parents react when the *yutori* reforms were implemented?
6. In your opinion, to what extent, did the *yutori* reforms achieved its intended goals?
7. Suppose you were on the *yutori* reforms advisory committee in 1998, what would you suggest?

Appendix B

Invitation Letter to Participate in Interview

Dear [Researcher],

My name is Meiko Lin. I am a doctoral candidate in the Interdisciplinary Studies Program at Teachers College, Columbia University in New York City. I am writing to invite you to participate in my doctoral dissertation entitled, *Examining the Influence of Yutori Education Reforms in Japan on Opportunity to Learn and Student Achievement on the TIMSS: A Multiple Cohort Analysis*. The aim of this study is to explore the effects of *yutori reforms* on opportunity to learn (OTL), and examine how the changes in OTL may have subsequently affected Japanese 8th graders' mathematics achievement as measured by the TIMSS data. This study hopes to extend theoretical understandings of the influences of the *yutori reforms* through examination of OTL. My invitation to you is based on my literature review on the *yutroi reforms* and nominations provided by Dr. Pizmony-Levy, Assistant Professor of International and Comparative Education at Teachers College. Collectively, we seek out individuals who have conducted extensive research studies on the *yutori reforms*.

This study uses a sequential, exploratory mixed methods design where qualitative data is explored first, followed by collection and analysis quantitative data. In the qualitative phase, I plan to gather interview data and archival data to get a clear picture of how *yutori reforms* were actually implemented in classrooms and schools. I hope to get a sense of which factors might need to be examined jointly with OTL during quantitative analysis through the qualitative phase.

Your participation in this study involves an interview lasting approximately one to two hours. The interview can be conducted face-to-face or online according to personal preferences. If needed, I may contact you within a three-month period with follow-upquestions, depending on your availability.

Your privacy is very important to me. I will treat the interviews, notes, and any other documents you provide with the utmost confidentially, and only I will have access to your identity. Your identity will be confidential, and will not be released to any persons in your institution or beyond it. Pseudonyms and other identity-masking techniques will be used in all presentations or writings about the study.

I hope that you will be able to join me in this study, and thereby contribute to an improved understanding of the influences of the *yutori reforms*. If you have any questions regarding the study, you can contact me via email at ml2734@tc.columbia.edu. I look forward to talking with you and thank you very much for taking the time to consider participation.

Sincerely,

Meiko