THREE ESSAYS ON TEACHER EDUCATION PROGRAMS AND TEST-TAKERS' RESPONSE TIMES ON TEST ITEMS

By

Hong Qian

A DISSERTATION

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

Measurement and Quantitative Methods- Doctor of Philosophy
Curriculum, Instruction, and Teacher Education- Doctor of Philosophy

2013

ABSTRACT

THREE ESSAYS ON TEACHER EDUCATION PROGRAMS AND TEST-TAKERS' RESPONSE TIMES ON TEST ITEMS

By

Hong Qian

This dissertation includes three essays: one essay focuses on the effect of teacher education programs on teacher knowledge while the other two focus on test-takers' response times on test items.

Essay One addresses the problem of how opportunities to learn in teacher education programs influence future elementary mathematics teachers' knowledge. This essay used data collected for the Teacher Education and Development Study in Mathematics (TEDS-M). TEDS-M measured the mathematics content knowledge (MCK) and the mathematics pedagogical content knowledge (MPCK) of future teachers in their final year in teacher preparation programs. The purpose of this essay is to explore whether elementary teaching candidates' MCK and MPCK are associated with their opportunities to learn in mathematics courses, mathematics methods courses, general pedagogy courses, and student teaching in five countries. The results showed that opportunities to learn in some teacher preparation components are more important than in other components and that there are more associations in high-performance countries than in low-performance countries.

Essay Two addresses the problem of how to detect item pre-knowledge using item responses and response time data. Item-knowledge is indicated by an unexpected short response time and unexpected correct response. This essay used a hierarchical framework proposed by other researchers for predicting expected response and response times. Large residuals between the

expected responses and the observed responses indicate aberrance. Two samples are used for detecting item pre-knowledge. The first sample is from the early stage of the operational test and is used for item calibration. The second sample is from the late stage of the operational test, which may feature item pre-knowledge. The purpose of this essay is to explore whether there are item pre-knowledge and compromised items in the second sample using the parameters estimated from the first sample. The results showed two items (out of 111) potentially exposed, and two candidates (out of 1,172) showing some indication of pre-knowledge on multiple items.

Essay Three addresses the problem of how to improve ability estimation or shorten a test using response times as collateral information. Response times can provide useful information about ability based on the correlation between person ability and person speed. The purpose of this essay is to explore whether incorporating a response time model into a CAT procedure will improve the correlation between true ability and estimated ability and classification accuracy by simulation studies. At the same time, the test length for the examinees whose ability is near the cut score is usually very long in variable length adaptive tests that have stopping rules related to decision making. Another purpose of this essay is to investigate whether using response times as collateral information can shorten the test length for these examinees while maintaining the same level of accuracy. The results from this study showed that using response times as collateral information does not improve ability estimation for this large-scale licensure examination using the specified estimation procedure which was designed to support a CAT procedure.

ACKNOWLEDGEMENTS

This three-essay dissertation was a product of generous support I received from Dr. Peter Youngs and Dr. Mark Reckase. I am incredibly fortunate to have two exceptional advisors. I am grateful for Dr. Peter Youngs' support on essay 1, the opportunity to work on several research projects and manuscripts with him, and his support of my pursuit of doctoral studies in Measurement and Quantitative Methods. I am also grateful for the support of Dr. Mark Reckase on essay 2 and essay 3. I appreciated the opportunity to work with him on several independent studies, which strengthened my knowledge of equating, computerized adaptive testing, standard setting, and cognitive diagnostic assessment. I will never forget the encouragement I received from my two advisors during my doctoral program, which gave me a lot of confidence as an international graduate student.

I want to thank Dr. Maria Teresa Tatto for permitting me to use the TEDS-M data for essay 1. For essay 2, I would like to thank Dr. David Scrams and Dr. Dorota Staniewska at the Financial Industry Regulatory Authority. They not only offered me a productive internship experience, but they also helped me grow as a researcher in the psychometrics field. Finally, I am indebted to Dr. Phil Dickison and Dr. Ada Woo at the National Council of State Boards of Nursing. They provided me with the data for essay 3, funding that enabled me to finish my dissertation, and finally the job I have now as a psychometrician.

Finally, it's hard to know how to express my appreciation for my mom, mother-in-law, my husband Li, and my two kids, April and Alan. I would not have finished my dissertation without their support. Last but not least, I want to thank Amy, my third child who is still in my belly, for being a good baby and letting me finish the dissertation at 36 weeks of pregnancy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. The effect of teacher education programs on future elementary mathematics teachers' knowledge with TEDS-M data

Numerous studies have found that teacher quality is the most important school-related factor affecting student achievement (Goe, 2007; Kaplan & Owings, 2001; Rice, 2003). Teachers' knowledge and classroom practices are consequential for students' learning (NCTAF, 1996). For example, teachers' knowledge influences the mathematics achievement of their students (Baumert et al., 2010; Hill, Rowan, & Ball, 2005). However, there is a large amount of variability among future mathematics teachers with regard to their knowledge. The Teacher Education and Development Study in Mathematics (TEDS-M), the first cross-national study about mathematics teacher education with large-scale samples, revealed large cross-country and within-country variation in the levels of future teachers' knowledge. Figure 1.1 in Appendix 1.1 shows the 16 within-country distributions of average future teacher performance on the mathematics content knowledge assessment for each sampled institution at the elementary level (Schmidt, Cogan, & Houang, 2011).[1]

As noted in the TEDS-M main report, "It is natural to wonder what accounts for differences in knowledge across and within countries. The answer to this question requires additional analyses, and is beyond the scope of this report" (Tatto et al., 2012, p 149). Consequently, my study is based on additional analyses that explore the effect of teacher

[1] There are 17 countries in the TEDS-M study. Figure 1 does not include Canada because in Canada, education is the responsibility of each province and TEDS-M was only conducted in four Canadian jurisdictions. At the same time, Figure 1 includes the data from U.S. private institutions, which are not part of the TEDS-M study.

preparation programs on future elementary mathematics teachers' knowledge using TEDS-M data.

In the first section of this essay, I review the research literature on teacher preparation and teacher knowledge. The second section describes my theoretical framework, which posits several associations between teaching candidates' preparation experiences and their knowledge. The third section introduces my research methods, including samples, instruments, measures, and analytical strategies.

Literature review

In this section, I review research on teacher preparation and teacher knowledge. Many studies have examined the relationship between teacher preparation and student learning by focusing on observed teacher characteristics, coursework, and features of student teaching. Recent research on teacher characteristics has found that teachers' certification status, performance on state certification tests, and attendance at a competitive undergraduate institution are all related to student learning gains in mathematics (Clotfelter, Ladd, & Vigdor, 2007; Croninger et al., 2007; Goldhaber & Brewer, 2000; Rowan, Chiang, & Miller, 1997; Wayne & Youngs, 2003).

In a review of research from the U.S., Clift and Brady (2005) reported indeterminate effects of mathematics methods courses and student teaching experiences on elementary mathematics teaching candidates. On one hand, studies by Langrall et al. (1996), Kim and Sharp (2000), and Mewborn (1999) found that methods courses and field experiences seemed to affect candidates' beliefs about teaching mathematics and their ability to demonstrate knowledge of constructivist principles in planning instruction. On the other hand, Vacc and Bright (1999) reported that despite changes in their beliefs about teaching mathematics, prospective elementary teachers were limited in their ability to employ knowledge of students' mathematicsematical thinking in their planning and instruction.

According to Schmidt, Bloemeke, and Tatto (2011), the inconsistent relationship between teacher education and prospective teacher competence is due to the fact that only crude indicators have been used as measures of opportunity to learn (OTL) in teacher preparation programs. The following excerpt from Schmidt et al. (2011) summarizes and critiques the findings from previous studies of teacher preparation:

> Many studies use the number of courses taken or the kind of teaching license (earned) to define OTL. So, not surprisingly, findings about the effects of content in teacher education on professional competence are inconsistent (Bloemeke, 2004; Cochran-Smith & Zeichner, 2005; Wilson, Floden, & Ferrini-Mundy, 2001) … It points to the need for more sophisticated measures of OTL than are presently available. Regardless of how common it is to use indicators like degrees, majors, examination results, or the number of classes taken (see, e.g., Akiba, LeTendre, & Scribner, 2007; Goldhaber & Brewer, 2000; Monk & King, 1994), this approach is at high risk of washing out any kind of relationship between opportunities to learn in teacher education and the outcomes because there is unfortunately nothing in teacher education "that share[s] a relatively common meaning across various cultural contexts" (Akiba, LeTendre, & Scribner, 2007, pp.65-66).

Recently, researchers have begun to (a) develop less aggregated measures that capture the content of teacher education in a low-inference way (Schmidt et al., 2011) and (b) examine associations between specific components of teacher preparation and student achievement. A U.S. study of New York City teachers by Boyd and colleagues (2009) examined the effects of preparation experiences on elementary teachers' effectiveness in teaching mathematics. The study found that several aspects of student teaching were associated with student learning in mathematics: whether or not the teachers had student taught, having had a supervisor who provided oversight of their student teaching, and the degree of alignment between their current school context and the school context in which they student taught. In addition, completion of mathematics methods courses was associated with effectiveness among second-year teachers, but not among first-year teachers (Boyd et al., 2009). This study used student achievement as a dependent variable and did not include measures of teacher knowledge.

Two recent international studies by Schmidt, Bloemeke, Tatto, and colleagues represent an important advance in research on pre-service mathematics teacher education and teacher knowledge. In the Mathematics Teaching in the 21st Century Study (MT21), Schmidt, Bloemeke, and Tatto (2011) investigated antecedents of mathematics content knowledge (MCK) and mathematics pedagogical content knowledge (MPCK) for lower secondary mathematics teachers in six countries. First, they reported that opportunities to study calculus and advanced mathematics were related to MCK in five areas of mathematics: number, geometry, algebra, function, and data. Second, they found that MPCK was associated with three types of practical experiences: the degree of opportunity related to instructional interactions around mathematics, the number of different types of practical experiences, and the number of weeks during which they had major responsibility for mathematics instruction during student teaching (Schmidt, Bloemeke, & Tatto, 2011). This study concentrated on the preparation of lower secondary mathematics teachers.

In the second study, Schmidt, Cogan, and Houang (2011) used TEDS-M data to examine country-level correlations between candidates' MCK and the average number of courses taken in mathematics and mathematics methods for both elementary and lower secondary mathematics teachers. At the elementary level, they reported that teaching candidates in the countries with the highest scaled scores of MCK took a significantly higher number of mathematics courses than their counterparts in lower-performing countries. In addition, Schmidt, Cogan, and Houang (2011) found that lower secondary candidates in countries with the highest MCK scores took almost twice as many mathematics courses and significantly more mathematics methods courses than their counterparts in lower-performing countries. This study focused on teacher subject-matter knowledge.

Taken together, the findings from these two studies indicate that for lower secondary mathematics candidates, MCK is associated with the number and nature of mathematics content courses taken while MPCK is associated with more practical experiences during teacher preparation. In addition, while the TEDS-M analysis conducted by Schmidt, Cogan, and Houang (2011) addressed predictors of elementary candidates' MCK, neither study investigated associations between elementary candidates' MPCK and their experiences in teacher preparation programs. My study is based on the work of Schmidt, Bloemeke, Tatto, and colleagues, and advances their work in three ways.

First, my study explores the effect of teacher preparation programs on elementary mathematics teachers' MCK and MPCK. Second, my study addresses predictors of elementary candidates' knowledge that are not included in the study by Schmidt, Cogan, and Houang (2011), such as opportunities to learn in mathematics education courses, general pedagogy courses and student teaching. Third, in the study by Schmidt, Cogan, and Houang (2011), the prior mathematics knowledge of these future teachers is only accounted for with a measure of whether they attend public or private institutions. In my study, I use more refined variables to control for their prior mathematics knowledge in order to account for differences in their knowledge before they enter teacher preparation.

Theoretical framework

In this section, I present the theoretical framework that undergirds my research design and data analysis. Teachers' professional knowledge includes knowledge of content, knowledge of pedagogy, and knowledge of content-specific pedagogy (Schmidt, Bloemeke, & Tatto, 2011;

Shulman, 1987). Among all types of teacher knowledge, MCK and MPCK are essential for future mathematics teachers to be effective (Allen, 2003; Goldhaber & Brewer, 1997; Hill et al., 2007; Ingvarson, Beavis, & Kleinhenz, 2007; Shulman, 1987). Teacher knowledge in my study is measured by these two types of knowledge.

In terms of research on teacher preparation, many scholars have used indicators such as certification status, certification test performance, degrees, and majors to draw inferences about the effects of university courses and student teaching on various teacher outcomes (Clotfelter, Ladd, & Vigdor, 2007; Croninger et al., 2007; Goldhaber & Brewer, 2000; Wayne & Youngs, 2003). But a growing number of researchers have criticized the use of such indicators because (a) they are at best approximations of candidates' experiences during teacher preparation and (b) they often lack common meaning within or across various national contexts (Akiba, LeTendre, & Scribner, 2007; Schmidt, Bloemeke, & Tatto, 2011). Instead, Schmidt, Bloemeke, Tatto, and colleagues have argued that research on teacher education should directly examine candidates' opportunities to learn in courses and field experiences.

In their research, Schmidt and colleagues define opportunity to learn as "the content to which future teachers are exposed as a part of their teacher preparation programs" (Schmidt, Cogan, & Houang, 2011, p.140). In addition, they differentiate the content to which candidates are exposed in mathematics teacher education into four main categories: mathematics, mathematics pedagogy, general pedagogy, and practical experiences (Schmidt, Bloemeke, & Tatto, 2011). My theoretical framework draws directly on the work of Schmidt and colleagues to posit that elementary candidates' opportunities to learn in (a) mathematics courses, (b) mathematics methods courses, (c) general pedagogy courses, and (d) student teaching are likely to be associated with their knowledge.

Mathematics courses

At the elementary level, Hill (2010) reported that teachers who took additional mathematics courses had higher levels of mathematicsematical knowledge for teaching (MKT). MKT includes common mathematics content knowledge that is held by a well-educated adult (such as an ability to calculate correctly), and specialized mathematics content knowledge that is unique to teachers (such as ability to respond to students' nonstandard mathematicsematical solutions) (Hill et al., 2007), which is similar to a combination of MCK and MPCK. According to Hill, "(a)lthough the correlation with mathematics content courses is significant, it is not substantively large" (2010, p.531). In addition, in her study Hill did not indicate whether completion of specific mathematics courses was associated with higher levels of MKT. At the lower secondary level, Schmidt, Bloemeke and Tatto (2011) reported from the MT21 study that opportunities to study calculus and advanced mathematics were significantly related to MCK in five areas: number, algebra, geometry, function, and data. But such opportunities did not have an effect on teachers' MPCK for teaching lower secondary mathematics. Taken together, the findings from these studies suggest the need to explore how the number and nature of the mathematics content courses taken by elementary teaching candidates are related to their knowledge. Therefore, in my study, I will use TEDS-M data to examine whether candidates took university-level courses in 19 areas of mathematics. This will enable me to test the following hypotheses:

Hypothesis 1A: *The number of the university-level mathematics content courses taken by a teaching candidate is positively related to their level of knowledge.*

Hypothesis 1B: *The nature of the university-level mathematics content courses taken by a teaching candidate is positively related to their level of knowledge.*

Mathematics methods courses

In a study in New York City, Boyd and colleagues (2009) investigated how variations in teacher preparation were associated with differences in mathematics and reading achievement for their students. They found that completion of a mathematics methods course was associated with effectiveness among second-year teachers. The study by Boyd et al. (2009) suggests that the central factors that influence beginning teachers' effectiveness in mathematics may not be the number of mathematics methods courses they have taken. Instead, the nature of their learning activities in these courses may be more consequential for student learning. In my study, I use TEDS-M data to test hypotheses about how elementary candidates' experiences in mathematics methods courses are associated with their knowledge:

Hypothesis 2A: *The number of topics addressed in mathematics methods courses taken by a teaching candidate is positively related to their level of knowledge.*

Hypothesis 2B: *The opportunities to learn (OTL) in mathematics methods courses for a teaching candidate are positively related to their level of knowledge.*

General pedagogy courses

In terms of general pedagogy courses, recent theoretical work by Bloemeke and colleagues has posited that general pedagogical knowledge is essential to effective teaching (Bloemeke et al., 2008; Konig et al., 2011). General pedagogical knowledge includes, but is not limited to, knowledge of instructional planning, student assessment, classroom management, how to facilitate group work, and how to address heterogeneity among students in instruction (Bloemeke et al., 2008; Shulman, 1987). Prior scholarship on general pedagogical knowledge suggests the need to explore how the number and the opportunities to learn in general pedagogy courses taken by elementary teaching candidates are related to their knowledge. I will use TEDS-M data to test the following hypotheses:

Hypothesis 3A: *The number of the topics addressed in general pedagogy courses taken by a teaching candidate is positively related to their level of knowledge.*

Hypothesis 3B: *The opportunities to learn (OTL) in general pedagogy courses for a teaching candidate are positively related to their level of knowledge.*

Practical experiences

In lower secondary mathematics, Schmidt, Bloemeke, and Tatto (2011) reported from the MT21 study that several practical experiences during teacher preparation were associated with candidates' MPCK. These practical experiences included the extent of opportunity related to mathematics-oriented teaching interactions, the number of distinct practical experiences, and the number of weeks during student teaching when they had primary responsibility for mathematics instruction. In addition, while the Boyd et al. (2009) study did not include measures of teacher knowledge, the researchers did report that elementary teachers' effectiveness was associated with having completed student teaching. Taken together, the findings from these studies suggest the need to explore how teacher knowledge is associated with the amount of time that candidates engage in student teaching with full responsibility for instruction and their opportunities to learn in student teaching. Therefore, in my study, I will use TEDS-M data to test the following hypotheses:

Hypothesis 4A: *The extent to which a teaching candidate engages in student teaching with full responsibility for instruction is positively related to their level of knowledge.*

Hypothesis 4B: *The opportunities to learn (OTL) in student teaching for a teaching candidate are positively related to their level of knowledge.*

In sum, my theoretical framework posits that elementary teaching candidates' knowledge (including MCK and MPCK) will be associated with their opportunities to learn in mathematics courses, mathematics methods courses, general pedagogy courses, and student teaching.

Method

Samples

TEDS-M was an international comparative study of teacher education that focused on the preparation of mathematics teachers at the primary (elementary) and lower secondary (middle school) levels in 17 countries (Tatto et al., 2012). The future elementary and middle school mathematics teachers in the study were in their final year in their teacher preparation programs. My study focuses on future elementary mathematics teachers. There are four different program types across and within countries that prepare future elementary teachers: lower-primary generalists (Grade 4 maximum); primary generalists (Grade 6 maximum); primary/lower-secondary generalists (Grade 10 maximum); and primary mathematics specialists.[2] Since most future teachers planning to work in primary schools are prepared as generalists who will teach classes no higher than Grade 6, my study focuses on programs who prepare primary generalists (Grade 6 maximum). There are six countries preparing primary generalists: Chinese Taipei, the Philippines, Singapore, Spain, Switzerland, and the U.S. The sample size for programs and future teachers for this group are in Table 1.1 in Appendix 1.2.[3] The TEDS-M sampling plan was developed to obtain nationally representative data for participating countries (for more

[2] The upper grade limit for "primary mathematics specialists" depends on how each country defines the grade level for primary school. For example, if a country defines grade 1-6 as primary school, then primary mathematics specialists can teach up to grade 6.

[3] Because the exclusion rate was greater than five percent for the Philippines, I did not include the Philippines for analysis.

details on sampling, see Tatto et al., 2012). I used pairwise deletion to delete the missing data for analysis.

Context

In this section, I provide some contextual information about the status of the teaching profession in each of the five countries, which may influence teaching candidates' MCK and MPCK and shape their experiences in teacher preparation programs. The contextual information included below was based on the *TEDS-M Encyclopedia: A Guide to Teacher Education Context, Structure and Quality Assurance in the Seventeen TEDS-M Countries* (Schwille, Ingvarson, & Holdgreve-Resendez, 2012).

Chinese Taipei. Teaching in traditional Chinese culture was high in prestige and status. Current Taiwanese political, economic, and social contexts have reinforced these advantages with generous salaries and other benefits for teachers. All remuneration for public school teachers is government funded, which is stable and ensured. Teaching is generally infused with dignity and authority, and therefore the role of teachers is viewed as similar to that of parents. Teacher schedules also make teaching attractive. Teachers are given time to perform many routine tasks in addition to teaching students. Job security and supportive work conditions also foster the appeal of teaching.

These characteristics make teaching very attractive, especially to people who seek stability in life. As a result, teaching in Taiwan is very competitive. Candidates face rigorous evaluations and serious competition throughout the teacher preparation process. This includes being selected to enter a preparation program, taking a teacher qualification examination, and job screening. The government sets the guidelines for recruiting, training and qualifying pre-service teachers, and developing in-service teachers' professional competencies.

Singapore. Singapore has a concerted set of national policies designed to ensure that teaching is an attractive career choice. Singapore is able to recruit teachers from the top one-third of each age cohort with regard to academic ability. In order to continue to attract a steady stream of talented and dedicated young adults into the Education Service within a competitive labor market, the remuneration package for teachers is pegged at a level that attracts a fair share of talent into the teaching profession compared to other professions.

The city-state of Singapore has only one teacher education institution, the National Institute of Education (NIE), which is an autonomous institute of Nanyang Technological University. As a result, the institution has maintained a high degree of control over teacher training and certification in the nation. Teachers are recruited by the Ministry of Education and sent to NIE for training.

Spain. Teachers in public schools in Spain are civil servants. Much of the teacher education curriculum for all universities is mandated by government-issued guidelines. This has been the case since the education system was created in the 18th century. Multiple laws and royal decrees continue to define and develop the complex framework of this system.

Initial salaries for teachers are relatively high. Although there are few career development and advancement opportunities, teachers can expect salary increases essentially based on length of service and undertaking professional development courses. The pension scheme is linked to a percentage of the base salary for civil servant teachers. The employment situation in Spain means that a position in a state primary educational institution is attractive in regions where there is a scarcity of professional work but less attractive in industrial areas. Furthermore, the characteristics of the career advancement system and the low social esteem of teaching, among other factors, make the teaching profession seem rather unattractive, in the opinions of the teachers themselves.

Switzerland. Switzerland's teacher education system has changed in fundamental ways in the last two decades, moving toward the integration of teacher education in higher education—a process that occurred in other countries long before this. At the same time, the Swiss have reduced, but by no means eliminated, important differences between cantons (regions). Although Switzerland has an open entry policy (every student who has successfully passed the Matura, the high-school exit examinations, has a legal right to enroll in a university), the academic requirements for graduation from secondary schools are relatively high.

Students entering teacher education programs are described as typically high achievers (e.g., in the top 20 percent of their age group). There is a strong interest in teacher education in Switzerland among talented high school and university graduates. Salaries for beginning teachers are comparable to other graduate salaries. Salaries for Swiss teachers are among the highest in the OECD countries. However, Swiss research shows a marked loss of attractiveness of the teaching profession in recent years. Fewer parents now advise their children to enter the profession. State efforts to reduce costs in recent years have led to reductions in actual earnings for teachers. Although starting salaries are attractive, they reach a plateau at a comparatively early point in the career. Opportunities for promotion as a teacher are limited unless teachers apply for a limited number of management positions and move out of teaching. The rate at which teachers are leaving their profession has increased.

United States. The U.S. has gradually shifted from local control toward centralization of teacher licensure or certification policy at the state and (to a lesser extent) the national level. At the same time, teacher education program-types, licensure requirements, and program accreditation requirements for primary school and lower secondary mathematics teaching have continued to vary significantly both within and across states.

Teachers in the U.S. usually have lower status and are paid less than other comparable careers. The NCES' Baccalaureate and Beyond (B&B) Study of over one million 1992-93 bachelor's degree recipients can be used to assess the competiveness of teaching with other professions. The teacher education major was considered a "career-oriented" major. Among other "career-oriented" major graduates, teachers' salaries lagged behind. In 2003, teachers earned an average of $43,800, as compared to salary ranges between $59,300 and $74,900 for graduates who had earned a career-oriented degree in engineering, computer science, health, or business. They had also started out with the lowest entry-level salary of the "career-oriented" group: $26,600 compared to the average of $32,700. While most other "career-oriented" degree holders earned more than graduates with "academic-oriented" majors (i.e., arts and humanities), teachers actually earned less than graduates from the "academic-oriented" group.

Instruments

TEDS-M developed three instruments: a future teacher questionnaire, an educator questionnaire, and an institutional program questionnaire. The future teacher questionnaire has four parts. The focus of each part and the time allotted to completing it are shown in Table 1.2.

Parts A, B, and D featured Likert-type scale items while part C consisted of two tests that assessed future teachers' knowledge. TEDS-M employed a rotated block design in order to measure the desired breadth and depth of knowledge. There were five primary booklets and each future teacher answered questions in two booklets.

An educator questionnaire was administered to teacher educators who instructed future teachers in the fields of (a) mathematics and mathematics pedagogy and (b) general pedagogy.

The questionnaire included Likert-type scale items asking about candidates' general academic background, teaching background, professional experience, research experience, field-based instruction, opportunities to learn in their courses, and beliefs about mathematics. The questionnaire also asked about candidates' perceptions of the degree of coherence and the effectiveness of their teacher preparation programs.

An institutional program questionnaire asked about descriptive program information, future teachers' backgrounds, selection policies, program content, field experiences, program accountability and standards, staffing, program resources, and reflections on the program.

Measures

I discuss here the variables employed in my analyses: the dependent variables (MCK score and MPCK score), the independent variables (number and nature of university-level mathematics content courses, topics addressed and OTL in mathematics pedagogy courses, topics addressed and OTL in general pedagogy courses, length of and OTL in student teaching), and the control variables (previous mathematics level and parents' education).

MCK score. One of the dependent variables, *MCK score*, was constructed from responses to the mathematics content knowledge assessment. This assessment was designed to measure advanced mathematics knowledge related to the appropriate school mathematics taught at the elementary level rather than the level of knowledge associated with advanced undergraduate academic mathematics courses such as the theory of complex functions (Bloemeke et al., 2011). The MCK assessment consisted of 74 items spanning four content subdomains: number and operations, algebra and functions, geometry and measurement, and data and chance. These were derived from the subdomains used in the assessment frameworks for IEA's Trends in Mathematics and Science Study (TIMSS). Rasch scaling was used to create individual-scaled scores for each future teacher. The international mean for the MCK scale was 500, and the standard deviation was 100.

MPCK score. Another dependent variable, MPCK score, was constructed from responses to the mathematics pedagogical content knowledge assessment which included three subdomains: curricular knowledge, planning for teaching and learning, and enacting teaching and learning. The assessment included 32 MPCK items, which had one of three formats: multiple-choice (MC), complex multiple-choice (CMC), or constructed response (CR). MPCK score was also reported in scaled scores for each future teacher generated through the use of item response theory, with a mean of 500 and standard deviation of 100.

Number of mathematics content courses. This variable is defined both at the individual level and the institutional level. At the individual level, this variable measures the number of university-level mathematics content courses taken by elementary teaching candidates. The future teacher questionnaire asked whether future teachers have studied 19 distinct mathematics topics. At the institutional level, this variable measures the number of university-level mathematics content courses required by the program. This question was included in the institutional program questionnaire. Including this variable at both levels enables me to explore both the effect of personal choice – to the extent that this is possible, as some students elected to take certain types of courses – and the effects of institutional policies, as some institutions require all students to take certain types of courses while others do not (Schmidt, Bloemeke, & Tatto, 2011).

Nature of mathematics content courses. In the TEDS-M study, the 19 distinct mathematics content courses or topics can be conceptually grouped into four broader categories representing university-level mathematics: geometry, discrete structure & logic, continuity &

functions, and probability & statistics. It is possible that the overall number of mathematics content courses is not related to MCK and MPCK, but that the kind of courses the future teachers took is associated with their knowledge.

Topics addressed in mathematics methods courses. This variable is also defined both at the individual level and the institutional level. The future teacher questionnaire asked whether candidates addressed 8 topics in mathematics methods courses. For each candidate, I counted the number of different topics addressed in mathematics methods courses. The institutional program questionnaire asked about the number of mathematics pedagogy courses required for the program.

OTL in mathematics methods courses. Since measures of the number of courses taken cannot fully capture experiences within courses, I use this group of variables to closely examine elementary candidates' experiences in mathematics methods courses. The future teacher questionnaire items #5 and #6 asked candidates to indicate how frequently they engaged in several activities in mathematics pedagogy courses, such as analyzing examples of teaching, using pupils' misconceptions to plan instruction, and so on. Response categories included: "Never" (coded as 1), "Rarely" (2), "Occasionally" (3), and "Often" (4). TEDS-M reported seven scaled scores for OTL in mathematics methods courses based on these two items: class participation, class reading, solving problems, instructional practice, instructional planning, assessment uses and assessment practice. The reliability for these seven scales is 0.85, 0.83, 0.78, 0.89, 0.90, 0.91, and 0.87, respectively.

Topics addressed in general pedagogy courses. Similar to the above two variables, this variable is also defined both at the individual level and the institutional level. The future teacher questionnaire asked whether candidates addressed 8 topics in general pedagogy courses. For each candidate, I counted the number of different topics addressed in general pedagogy courses. The institutional program questionnaire asked about the number of general pedagogy courses required for the program.

OTL in general pedagogy courses. The future teacher questionnaire items #8 and #9 asked future teachers to indicate how frequently they engaged in several activities in general pedagogy courses, such as studying stages of child development, identifying appropriate resources needed for teaching, and so on. TEDS-M reported three scaled scores for OTL in general pedagogy courses based on these two items: teaching for diversity, teaching for reflection on practice, and teaching for improving practice. The reliability for these three scales is 0.89, 0.93, and 0.93, respectively.

Student teaching. The future teacher questionnaire asked what proportion of the time they had full responsibility for teaching their class during student teaching.

OTL in field experience. TEDS-M reported three scaled scores for OTL in field experience based on future teachers' responses to two items (Q13 and Q14): connecting classroom learning to practice, supervising teacher reinforcement of university goals for practicum, and supervising teacher feedback quality. The reliability for these three scales is 0.95, 0.94, and 0.95, respectively. The future teacher survey Q12 asked, "For about how much of the time in the field experience was one of your assigned supervisors present in the same room as you?" I also examine how this variable is associated with teaching candidates' MPCK.

Previous mathematics level. This control variable is at the individual level, and was constructed from responses to two questions on the future teacher questionnaire: (1) What was the highest grade level at which you studied mathematics in secondary school? And (2) In secondary school, what was the usual level of grades that you received?

Parents' education. This control variable is also at the individual level, and was constructed from responses to two questions on the future teacher questionnaire: (1) What is the highest level of education completed by your mother, and (2) What is the highest level of education completed by your father?

Analytical strategies

I used multilevel linear models to explore the effects of teacher preparation programs on future teachers' knowledge since future teachers were nested within preparation programs. The dependent variables MCK and MPCK are defined at the individual level while the independent variables are defined at both the individual and institutional levels. Variables defined at the individual level of the analysis highlight the effects of individual differences or choices while variables defined at the institutional level reflect differences between institutions in policies and practices (Schmidt, Bloemeke, & Tatto, 2011). A population model for MCK is:

Student level:

$$MCK_{ij} =\beta 0_{j}+ \beta 1_{j}(\text{Parent education}_{ij}) +\beta 2_{j}(\text{Previous mathematics level}_{ij})$$

$$+ \beta 3_{j}(\text{independent variable}_{ij})+u_{ij}$$

Program level:

$$\beta 0_{j} = \Upsilon 00+\Upsilon 01(\text{independent variable}_{j})+ v0_{j}$$

in which MCK_{ij} represents the MCK score for teaching candidate i in program j; independent variable_{ij} represents an independent variable at the individual level, such as number of topics addressed in mathematics methods courses taken by teaching candidate i in program j; independent variable j represents an independent variable at the institutional level, such as the number of mathematics methods courses required by program j which may have an effect on the intercept for the first level model.

Although these independent variables represent my central theoretical concerns, other factors could affect teaching candidates' MCK scores. Therefore, I include control variables such as candidates' previous mathematics level and parents' levels of education. Controlling for these measures helps us account for differences among teaching candidates that were manifest prior to entering teacher preparation. A similar model for MPCK score is:

Student level:

$$\text{MPCK}_{ij} = \beta 0_j + \beta 1_j(\text{Parent education }_{ij}) + \beta 2_j(\text{Previous mathematics level}_{ij})$$

$$+ \beta 3_j(\text{independent variable}_{ij}) + u_{ij}$$

Program level:

$$\beta 0_j = \Upsilon 00 + \Upsilon 01(\text{independent variable }_j) + v0_j$$

Results

Effect of number of mathematics content courses

Table 3 reports the standardized coefficients from a multilevel linear regression model for the five countries. The dependent variable is teaching candidate's MCK score. The independent variables are the number of mathematics content courses at the individual level and at the institutional level. The number of mathematics content courses at the individual level means the number of mathematics content courses the teaching candidate has taken while the number of mathematics content courses at the institutional level means the number of mathematics content courses that the institution required. The model also included two control variables: teaching candidate's previous mathematics achievement and parents' education.

The results in Table 1.3 showed that in two countries: Chinese Taipei and Switzerland, the coefficient for the number of mathematics content courses taken by teaching candidates is statistically significant, which indicates that there is an association between the number of mathematics content courses taken by teaching candidates and their MCK score even when controlling for the number of mathematics content courses required by the institution, teaching candidates' previous mathematics performance and their parents' education. In other words, the more mathematics content courses a teaching candidate in Chinese Taipei and Switzerland took, the higher his/her MCK score is, keeping the other three variables constant. But this is not the case in Singapore, Spain, and the U.S. For these three countries, the coefficient for the number of mathematics content courses taken by teaching candidates is not statistically significant.

The average number of mathematics content courses required by the institution in each country and standard deviation are as follows: Chinese Taipei: 3.24 (standard deviation is 4.428); Singapore: 3.04 (5.228); Spain: 0.60 (.799); Switzerland: 1.28 (1.264); and the U.S.: 2.61 (2.985). The effect of the number of mathematics content courses required by the institution is mixed, according to Table 3. In Singapore, the coefficient for the number of mathematics content courses required by the institution is positive, which indicates that the greater the number of mathematics content courses required by the institution, the higher the MCK scores for teaching candidates in this institution. However, in Spain and the U.S., the coefficient is negative, which means the greater the number of mathematics content courses an institution required, the lower the MCK scores for teaching candidates, when the other three variables were held constant. One possibility could be that if the previous mathematics performance for all teaching candidates in an institution is generally low, then the institution may require that they take a greater number of mathematics content courses. However, even when they take more required mathematics content courses, their mathematics content knowledge at the completion of the program is still low compared to students in other institutions with higher previous mathematics performance and who took a smaller number of required mathematics content courses.

While Table 3 focuses on MCK score, Table 4 focuses on MPCK score. In Chinese Taipei and Singapore, there is a positive association between the number of mathematics content courses teaching candidates took and their MPCK scores while in the U.S. there is a negative association between these two variables. For the institution-level variable, the effect is also mixed. In Singapore, the number of mathematics content courses required by the institution is positively associated with teaching candidates' MPCK score. However, in Spain, the greater the number of mathematics content courses required by the institution, the lower the MPCK scores for teaching candidates at this institution.

The results for the control variables from Table 1.3 and Table 1.4 are worth noting. Teaching candidates' previous mathematics achievement has a significant association with their MPK and MPCK across almost all countries (except Spain for MPCK). Therefore, it is important to include this variable to control for the difference in teaching candidates' mathematics ability before they enter teacher preparation. When it comes to parents' education, this variable does not have any effect on teaching candidates' MCK or MPCK, except for candidates in the U.S. In the U.S., there is a strong relationship between teaching candidates' parents' education levels and their MCK and MPCK. For MCK, the effect size of parent education is as large as teaching candidates' previous mathematics ability. This indicates that for U.S. teaching candidates, their MCK and MPCK are significantly associated with their parents' education levels, which means that teacher preparation programs cannot reduce the inequity produced by family background. This is not the case in the other four countries.

Since the overall number of mathematics content courses did not have a consistent impact on MCK score, I also examined whether taking particular mathematics content courses made any difference. In the TEDS-M study, the 19 distinct mathematics content courses or topics can be conceptually grouped into four broader categories representing university-level mathematics: geometry, discrete structure & logic, continuity & functions, and probability & statistics. I counted how many courses teaching candidates took in each category and used this as an independent variable. I included each of the independent variables and control variables (previous mathematics, parents' education, the number of mathematics content courses the institution required) in the model and estimated the effect of the number of mathematics content courses in a certain category on teaching candidates' MCK score. Table 1.5 reported the unstandardized coefficients for each of the four independent variables in five countries (the coefficient for control variables are not reported).

The results in Table 5 showed that for Chinese Taipei teaching candidates, if they took one more course in Geometry (range from 0-4), their MCK score would increase 5.85 compared to other candidates with the same previous mathematics score, the same level of parents' education and the same number of total mathematics content courses required by the institution. Note that the international mean for MCK score is 500. The effect is significant, which means that the increase is different from 0. However, for U.S. teaching candidates, if they took one more course in Geometry, their MCK score would be significantly lower by 6.04 than other candidates, keeping the other three variables constant. The effect of the number of mathematics content courses in discrete structure & logic, continuity & functions, and probability & statistics on MCK is more consistent (significant in at least two countries). When the total number of mathematics content courses required by the institution is the same, it seems that taking a greater number of mathematics courses in continuity & functions gives teaching candidates more knowledge. For Chinese Taipei teaching candidates, taking one more course in continuity and functions (range from 0-6) would lead to an increase of 14.22 in their MCK score, while for teaching candidates in Switzerland and U.S., the increase is 6.47 and 6.12, respectively.

Effect of number of topics addressed in mathematics education courses

The results in Table 1.6 showed that the coefficient for the number of topics addressed in mathematics education courses (at the individual student level) is positive and statistically significant at the 0.05 level in Chinese Taipei and the 0.001 level in Switzerland. This indicates that there is an association between the number of topics addressed in these courses and teaching candidates' MCK score. Teaching candidates in these two countries who addressed more topics in these courses were more likely to have higher levels of MCK. However, there is no positive effect for the number of mathematics education courses required by the institution (there is one negative coefficient in Switzerland). For MPCK, the results in Table 1.7 showed a similar pattern.

Similar to the nature of mathematics content courses, I also examined which topics addressed in mathematics education courses were most important. In the TEDS-M study, the 8 mathematics education topics can be conceptually grouped into two categories: foundations and instruction. Foundations include three topics, such as foundations of mathematics, context of mathematics education, and development of mathematics ability and thinking. Instruction includes five topics: mathematics instruction, developing teaching plans, mathematics teaching (observation, analysis and reflection), mathematics standards and curriculum, and affective issues in mathematics. I counted how many topics teaching candidates addressed in each category and used this as an independent variable for MCK and MPCK. All models included three control variables: previous mathematics, parents' education, and the number of mathematics education courses the institution required. Table 1.8 reported the coefficients in four models for five countries (the coefficient for control variables are not reported).

The results in Table 8 showed that the coefficient for the number of topics addressed in foundations is positive and statistically significant in Switzerland for both MCK and MPCK. However, the coefficient for Singapore is negative and statistically significant for both MCK and MPCK. On the other hand, the effect of the number of topics addressed in instruction is more consistent. There was a positive and statistically significant effect in three countries for both MCK and MPCK.

Since measures of the number of topics addressed cannot fully capture experiences within mathematics education courses, I used OTL in mathematics education courses to closely examine elementary candidates' experiences in mathematics methods courses. TEDS-M reported seven scaled scores for OTL in mathematics methods: class participation, class reading, solving problems, instructional practice, instructional planning, assessment uses and assessment practice. The items used to construct each scaled score are listed in Table 1.9.

I used the OTL variables as independent variables, and put them into the model one at a time. The control variables for each model are: previous mathematics, parents' education, and the number of topics learned by the candidate. The dependent variable is MPCK. Table 1.10 reported the coefficient for each OTL variable in each country (the coefficient for control variables are not reported).

The results in Table 1.10 showed that the coefficients for class participation, instructional practice, and instructional planning are positive and statistically significant in Chinese Taipei, which indicates that teaching candidates who have more opportunities to do these things are more likely to have higher MPCK scores even though they took the same number of topics in mathematics education courses. The coefficients for class reading, solving problems, and assessment practice are also positive and statistically significant in Chinese Taipei, but in other countries, the coefficients for these three variables are negative and statistically significant, which means the effect of these three OTL variables are not consistent across countries. Therefore, Chinese Taipei is the only country where candidates' experiences in mathematics methods classes were positively associated with MPCK. However, I did not find the same effects in other countries.

Effect of number of topics addressed in general pedagogical courses

The results in Table 1.11 showed that the coefficients for the number of topics addressed in general pedagogical courses (both at the individual level and at the institutional level) are positive and statistically significant in Singapore, for MCK and MPCK. In Chinese Taipei, the number of general pedagogical courses required by the institution is associated with the MPCK score.

Similar to the topics in mathematics education courses, I also examined which topics addressed in general pedagogical courses are most important. In the TEDS-M study, the 8 general pedagogical topics can be conceptually grouped into two categories: social science and application. Social science includes three topics: history of education and educational systems, philosophy of education, and sociology of education. Application includes five topics: educational psychology, theories of schooling, methods of educational research, assessment and measurement, and knowledge of teaching. I counted how many topics teaching candidates addressed in each category and used this as an independent variable for MCK and MPCK. All models included three control variables: previous mathematics, parents’ education, and the number of general pedagogical courses that the institution required. Table 1.12 reported the coefficients in four models for the five countries (the coefficients for control variables are not reported).

The results in Table 1.12 show that the coefficient for the number of topics in social science is positive and statistically significant in Singapore, Spain, and Switzerland for MCK. For MPCK, it is positive and statistically significant in Singapore and Spain. The coefficient for the number of topics in application is positive and statistically significant in Singapore for both MCK and MPCK. However, it is negative and statistically significant in the U.S.

I also examined the effect of OTL in general pedagogical courses on MPCK. TEDS-M reported three scaled scores for OTL in general pedagogical courses: teaching for diversity, teaching for reflection on practice, and teaching for improving practice. The items used to construct each scaled score are listed in Table 1.13.

I used OTL variables as independent variables, and put them into the model one at a time. The control variables for each model are: previous mathematics, parents' education, and the number of general pedagogical topics learned by the candidate. The dependent variable is MPCK. I found that all coefficients are not statistically significant, which indicates that the opportunities to learn in general pedagogical courses are not associated with MPCK score.

Effect of student teaching

I hypothesized that the extent to which a teaching candidate engages in student teaching with full responsibility for instruction is positively related to their level of knowledge. The results from the analysis did not support this hypothesis.

I also examined the effect of OTL in student teaching on MPCK. TEDS-M reported three scaled scores for OTL in student teaching: connecting classroom learning to practice, supervising teacher reinforcement of university goals for practice, and supervising teacher feedback quality. The items used to construct each scaled score are listed in Table 1.14.

Table 1.15 showed that the coefficient for supervising teacher feedback quality is positive and statistically significant in Singapore (the coefficients for the control variables are not reported). I also examined the association between MPCK and the amount of time when the supervisor was present in the same room as the teaching candidate during student teaching. The coefficient is positive and statistically significant in Chinese Taipei, Spain, and Switzerland, which means in these countries, the more time the supervisor is present in the classroom with the teaching candidate, the higher the teaching candidate's MPCK, after controlling for previous mathematics performance and parents' education. However, the coefficient is negative and statistically significant in Singapore.

Conclusion

The main purpose of this study was to investigate associations between teacher preparation components and teaching candidates' knowledge based on data from TEDS-M study. Research has found that elementary teachers' knowledge is associated with student learning in mathematics (Hill et al., 2007; Hill, Rowan, & Ball, 2005). But there is little understanding in the research literature of how elementary teaching candidates acquire MCK and MPCK in different countries. In this study, I hypothesized that a number of aspects of teacher education might be relevant to the acquisition of MCK and MPCK, such as the number and nature of university-level mathematics content courses, exposure to topics in mathematics methods courses and general pedagogy courses, proportion of student teaching with full responsibility for instruction, and opportunities to learn in coursework and student teaching. A summary of the statistically significantly positive effects of these independent variables on MCK and/or MPCK in five counties is presented in Table 1.16.

From Table 16, hypotheses 1A, 1B, 2A, 2B, 3A, and 4B are supported in at least one country while hypotheses 3B and 4A are not supported in any countries included in this study. It seems that opportunities to learn (OTL) in general pedagogy courses for a teaching candidate are not positively related to their level of MPCK. Also, the extent to which a teaching candidate engages in student teaching with full responsibility for instruction is not positively related to their level of MPCK.

From Table 1.16, we can also see that in some countries, there are more significantly positive relationships between teacher preparation components and teaching candidates' knowledge than in other countries[4]. For example, in Chinese Taipei, the relationships between teacher preparation components and knowledge are very persistent. However, in the U.S., there is only one significantly positive relationship between teacher preparation components and knowledge. This may relate to the levels of MCK and MPCK in these countries. Table 1.18 lists the mean score of MCK and MPCK for each country and the number of significantly positive relationships between teacher preparation components and knowledge. It seems that in general the countries with higher levels of MCK and MPCK have more significantly positive relationships between teacher preparation components and knowledge.

However, this result needs to be interpreted with caution since there are other possibilities associated with it. One possible explanation is related to selection effects associated with

[4] In order to exclude the possibility that this may related to the variance of MCK/MPCK in these countries, I have checked the relationships between the variance (standard deviation) of MCK/MPCK and the number of significant associations between teacher preparation components and MCK/MPCK. I found that there is no relationship between them. Table 1.17 listed the result.

entering teacher preparation programs. In some countries, such as Chinese Taipei, the teacher education programs are very selective, which makes teaching candidates' MCK and MPCK uniformly high and have less variance before they enter teacher preparation programs. However, in counties like US, the teaching candidates' MCK and MPCK can vary to a large degree before they enter teacher preparation programs. Even though this study included some control variables, they may not fully count for these differences, which may artificially make various teacher preparation components seem more effective in some countries and not in others. Another possibility is that the reliability of the MCK and MPCK measures is higher in high-performance countries than in low-performance countries, which could more easily make the regression effect significant.

Discussion

Many studies have investigated teacher preparation components, most focusing on observable characteristics, such as the selectivity of the program, the number of content courses, the number of methods courses, or the length of student teaching assignments. While such factors seem important for teaching candidates, what's more important is what actually happens in coursework and student teaching. In this study, I drew on subscales for OTL from the TEDS-M study to explore the relationship between teaching candidates' experiences during teacher preparation and their level of MCK and MPCK. By examining teaching candidates' opportunities to learn in teacher preparation programs, this study went beyond focusing on observable characteristics of teacher preparation programs, and started to look at the nature of learning activities that occur in coursework and student teaching.

In this study, I hypothesized that the number and nature of the university-level mathematics content courses taken by a teaching candidate would be directly related to their level of knowledge, especially MCK. Many studies have reported that teachers' content knowledge is critical to student learning. For example, Rowan, Chiang, and Miller (1997) found that students who were taught by a teacher with a bachelor's or master's degree in mathematics or one who had scored well on a brief mathematics quiz had higher gains in achievement in this subject area. Goldhaber and Brewer (1997) reported a greater influence on student achievement of teachers' bachelor's and master's degrees in the content area taught (e.g., mathematics or mathematics education) than was true for undifferentiated degrees. In a multi-level analysis of the Longitudinal Study of American Youth (LSAY) data set, Monk and King (1994) reported some evidence of cumulative effects of teachers' prior as well as proximate subject matter coursework on student performance in mathematics. Monk (1994) also found that teachers' content preparation, as measured by coursework in the subject field, was usually positively, though rarely significantly, related to student achievement in mathematics and science. Finally, Mullens, Murnane, and Willett (1996) found that third-grade students in Belize learned more mathematics when their teachers had a strong command of the subject.

Based on the findings from these studies, I expected that the number and nature of the university-level mathematics content courses taken would influence teaching candidates' level of knowledge, especially MCK. The study found that the number of mathematics content courses taken has an effect on teaching candidates' level of MCK in three countries and on teaching candidates' MPCK in two countries. In terms of the nature of mathematics content courses, taking more courses in particular areas, such as discrete structure & logic, and continuity & functions, has an effect on teaching candidates' MCK in three countries. This result suggests that what is important is not only how many mathematics content courses the program offers, but also what kinds of courses it offers. It is also worth noting is that I examined elementary teaching candidates' knowledge; it would be interesting to see whether the results differ when the focus is on teaching candidates' knowledge at the middle school or high school levels.

Many studies have found that the number of mathematics methods courses is related to the teaching candidates' effectiveness in their future teaching. Begle (1979), Boyd et al. (2009), Briscoe and Stout (1996), Clift and Brady (2005), Kim and Sharp (2000), Langrall, Thornton, Jones, and Malone (1996), Mewborn (1999), and Monk (1994) all found a positive impact of methods courses on teaching candidates' beliefs and practices. For example, using data from the National Longitudinal Study of Mathematicsematical Abilities, Begle (1979) reported that teachers' coursework in mathematics methods had a stronger effect on student achievement than additional higher-level coursework in mathematics for a group of already strong teachers. Monk (1994) found that coursework in teaching methods had a stronger influence than additional coursework in mathematics on student achievement in mathematics and science. And Boyd and colleagues (2009) reported that required mathematics methods courses were associated with effectiveness among second-year teachers, though not among first-year teachers. The findings suggest that the greater the number of mathematics methods topics that teaching candidates have an opportunity to learn, the higher the level of their MCK and MPCK.

My study confirmed that the number of topics encountered in mathematics methods courses was associated with MCK in two countries and MPCK in one country. In addition, my results indicated that candidates' exposure to topics in instruction (such as developing teaching plans) was associated with MCK and/or MPCK in four countries while candidates' exposure to topics in foundations (such as foundations of mathematics) was associated with MCK and MPCK only in Switzerland. Further, opportunities to learn in mathematics methods courses (except assessment uses) were related to MPCK in Chinese Taipei, but not in the other four countries. These findings are consistent with a key result from the MT21 study. In that study, Schmidt, Blömeke, and Tatto (2011) reported that lower secondary mathematics candidates' PCK for teaching mathematics was associated with their degree of opportunity related to instructional interactions around mathematics and the number of different types of practical experiences that they had.

In this study, I hypothesized that the extent to which a teaching candidate engages in student teaching with full responsibility for instruction would be positively related to their level of MPCK. The hypothesis was not supported in this study. This finding is not consistent with findings from other research, in which Youngs and I (Youngs & Qian, 2013) investigated the association between teacher preparation components and Chinese elementary teaching candidates' MKT.

An interesting finding from that study is that the extent to which a teaching candidate engages in student teaching with full responsibility for instruction was directly related to their level of MKT while the overall length of student teaching (including student teaching when they did not have full responsibility for instruction) did not seem to matter. One possible explanation for the inconsistency between that study and my finding might be the different measures for this variable in the two studies. In the Youngs and Qian study, the instrument directly asked about the number of weeks of student teaching that the respondent completed when they had full responsibility for instruction.

However, in the TEDS-M study, the instrument asked what proportion of the time the respondent had full responsibility for teaching their class during student teaching. Response categories included: "less than 1/4 of the time" (code as 1), "1/4 or more, but less than 1/2" (code as 2), "1/2 or more, but less than 3/4" (code as 3), and "3/4 or more" (code as 4). It is possible that for some teaching candidates, the proportion is high but the absolute length of student teaching is very short. Therefore, it is possible that this study cannot capture the effect of student teaching with full responsibility for instruction on teaching candidates' knowledge.

A puzzling finding from this study is that for U.S. teaching candidates, only the number of mathematics content courses in continuity & function was associated with their MCK. Other proposed relationships are not supported in the U.S. sample while in the other four countries, more hypotheses are supported. One possible reason is the effects of the control variables. In terms of the effect of previous mathematics achievement, there is a statistically significant association between prior mathematics achievement and teaching candidates' MCK and MPCK across all countries. It is not surprising that teaching candidates' mathematics achievement before they enter their teacher preparation program influences their MCK and MPCK while they are in teacher preparation. This study provided a strong evidence for this association.

When it comes to the effect of parents' education, it is interesting to see that it is only statistically significant for the U.S., but not for the other four countries. For the U.S., the effect of parents' education is consistent for all models. The effect size of parents' education is similar to the effect size of previous mathematics achievement (compare the standardized coefficients) for the U.S. teaching candidates for MCK; and the effect size of parents' education is about half of the effect size of previous mathematics achievement when the dependent variable is MPCK. It seems that for U.S. teaching candidates, their MCK and MPCK mostly depends on their previous mathematics achievement and their parents' education, and their experiences in teacher preparation have little effect on their MCK and MPCK. This result is parallel to the finding that K-12 student achievement is mostly related to students' family background rather than school factors. This study expands the finding into post-secondary level schooling. It is a somewhat disappointing finding for policy makers who are seeking to improve teaching candidates' knowledge by reforming teacher preparation programs.

Another surprising finding from this study is that while some expected positive relationships were found in some countries, the same relationships were negative and statistically significant in other countries. One possible explanation is that the terms such as "course" and "topic"may mean very different things in different countries; as Akiba, LeTendre, and Scribner (2007, pp.65-66) point out, there is unfortunately very little in teacher education "that share[s] a relatively common meaning across various cultural contexts." For example, a "course" may mean several weeks of instruction in one country while in another country it may last for a full year. Therefore, the inconsistent results do not necessarily mean that a given aspect of teacher preparation is not important. It merely points to the need for more sophisticated measures of OTL than are presently available in the TEDS-M data. The limitation of survey data in examining teaching candidates' experiences in coursework and student teaching in different countries calls for qualitative research that would involve interviewing candidates about their preparation program experiences in various cultural contexts. Furthermore, this study only used two forms of teacher knowledge as outcome variables. For future research, other outcome variables can be examined, such as mathematics instruction and student learning.

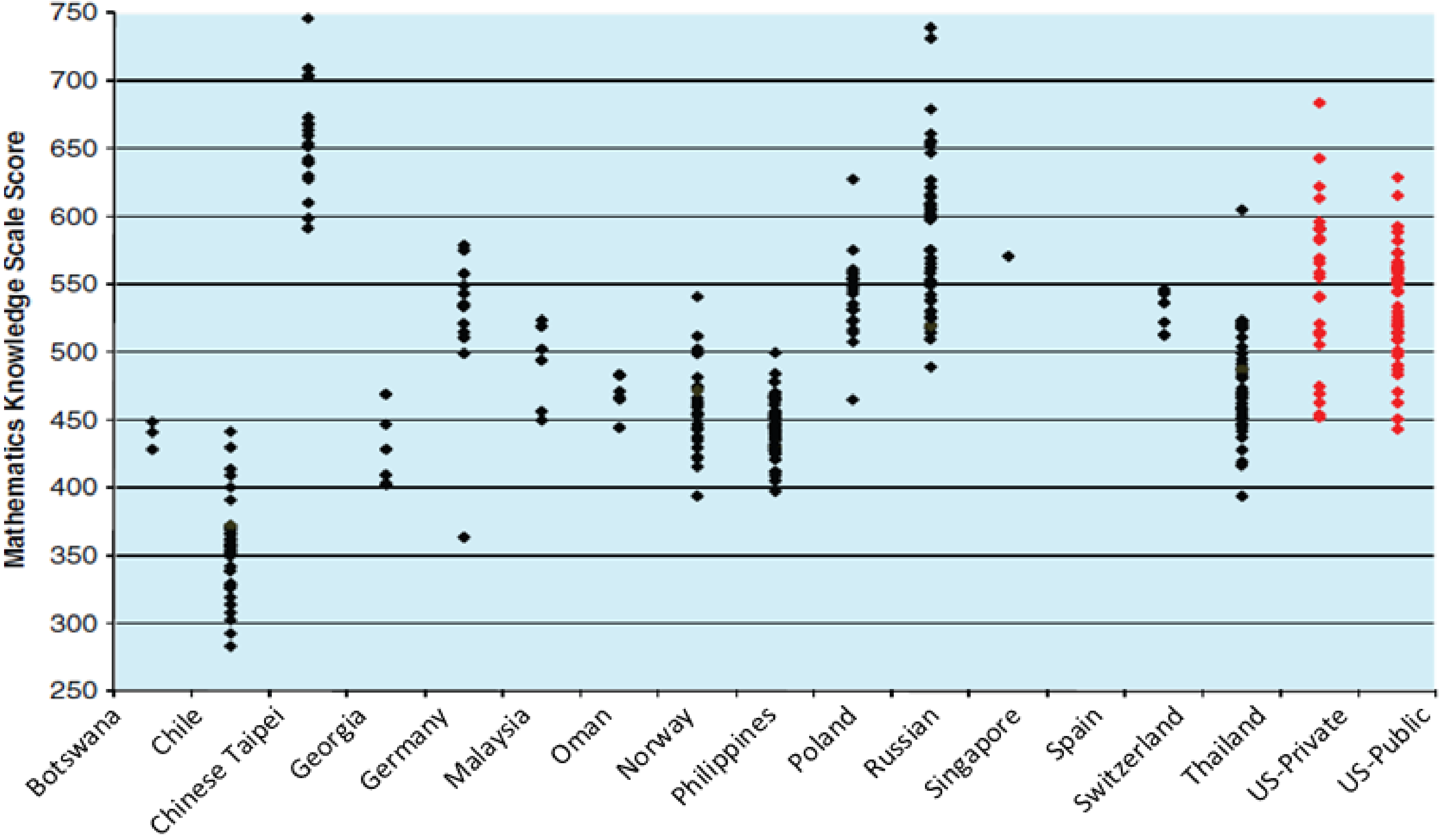
Even though there are limitations to this study, this research informs theory regarding opportunity to learn in teacher preparation (Schmidt, Blömeke, & Tatto (2011) and highlights how teacher preparation components can influence elementary teaching candidates' professional knowledge in five countries. Although there is a growing body of research on pre-service teacher education, more needs to be understood about how different aspects of teacher preparation can influence teachers thoseconsistent across several countries) can help researchers and policy makers to better' knowledge. Therefore, the findings from this study (especially understand the relative contributions of various program features to elementary teaching candidates' MCK and MPCK. This work can inform theory as well as future research regarding ways to structure and improve elementary mathematics teacher preparation by discerning how preparation experiences can shape teachers' knowledge. Such research has the potential to identify effective means for developing a valuable national asset – elementary mathematics teachers.

APPENDICES

Appendix 1.1

Figure 1.1

Institution-level mathematics knowledge scale scores by country at the elementary level [5]



[5] For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Appendix 1.2

Table 1.1

Sample sizes for preparation programs and future teachers in primary generalists program group

Country	Number of Preparation Programs	Number of Future Teachers
Chinese Taipei	11	923
Philippines	33	592
Singapore	6	263
Spain	48	1,093
Switzerland	21	815
United States	71	1,310
Total	190	4,996

Table 1.2

Composition of future teacher questionnaire

Section	Focus	Time (minutes)
A	General background	5
B	Opportunity to learn in teacher preparation program	15
C	MCK and MPCK	60
D	Beliefs about mathematics and teaching	10

Table 1.3

Estimated effects of number of mathematics content courses on MCK score

	Pre_mat	Par_edu	Mat_con_cou (candidate)	Mat_con_cou (institution)
Chinese Taipei	.30***	.01	.16***	-.02
Singapore	.22***	-.01	.10	.15*
Spain	.37***	-.20	.03	-.10*
Switzerland	.19***	.05	.07*	-.05
United States	.22***	.21***	-.04	-.12*

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

Table 1.4

Estimated effects of number of mathematics content courses on MPCK score

	Pre_mat	Par_edu	Mat_con_cou (candidate)	Mat_con_cou (institution)
Chinese Taipei	.25***	-.03	.11***	-.05
Singapore	.14**	-.01	.12*	.05*
Spain	.20	.01	-.03	-.10*
Switzerland	.12**	.03	.02	-.03
United States	.22***	.13***	-.07*	-.05

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

Table 1.5

Estimated effects of nature of mathematics content courses on MCK score

	Geometry	Discrete structures and logic	Continuity and functions	Probability and statistics
Chinese Taipei	5.85*	5.78**	14.22***	12.55**
Singapore	2.36	3.7	5.31	7.83
Spain	-.09	2.31**	.96	-2.82
Switzerland	.01	4.67**	6.47***	6.19*
United States	-6.04***	-1.78	6.12***	-1.77

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

Table 1.6

Estimated effects of number of topics addressed in mathematics education courses on MCK score

	Pre_mat	Par_edu	Mat_edu_top (candidate)	Mat_edu_cou (institution)
Chinese Taipei	.32***	.01	.06*	-.01
Singapore	.27***	.01	-.01	-.04
Spain	.37***	-.01	.01	.05
Switzerland	.21***	.05	.12***	-.09*
United States	.22***	.20***	-.01	.02

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

Table 1.7

Estimated effects of number of topics addressed in mathematics education courses on MPCK score

	Pre_mat	Par_edu	Mat_edu_top (candidate)	Mat_edu_cou (institution)
Chinese Taipei	.27***	-.03	.03	-.05
Singapore	.19***	.01	-.01	-.06
Spain	.19***	.03	.05	.01
Switzerland	.14***	.02	.11***	-.19***
United States	.21***	.13***	.03	-.01

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

Table 1.8

Estimated effects of nature of mathematics education topics on MCK and MPCK scores

	Foundation		Instruction	
	MCK	MPCK	MCK	MPCK
Chinese Taipei	2.25	-.13	4.09*	2.29
Singapore	-10.12*	-9.79**	6.62*	6.3*
Spain	-1.18	.03	1.3	3.91*
Switzerland	6.37**	6.2**	4.83**	4.10*
United States	-2.20	-.03	1.3	2.60

*p<0.05; **p<0.01; ***p<0.00

Table 1.9

Lists of items used to construct each OTL variable in mathematics education courses

OTL	In the mathematics education<pedagogy/teaching methods> courses that you have taken or are currently taking in your teacher preparation program, how frequently did you do any of the following? (Never, Rarely, Occasionally, Often)
Class participation	Ask questions during class time Participate in a whole class discussion Make presentations to the rest of the class Teach a class session using methods of my own choice Teach a class session using methods demonstrated by the instructor
Class reading	Read about research on mathematics Read about research on mathematics education Read about research on teaching and learning Analyze examples of teaching (e.g., film, video, transcript of lesson)
Solving problems	Write mathematicsematical proofs Solve problems in applied mathematics Solve a given mathematics problem using multiple strategies Use computers or calculators to solve mathematics problems
Instructional practice	Explore how to apply mathematics to real-world problems Explore mathematics as the source for real-world problems Learn how to explore multiple solution strategies with pupils Learn how to show why a mathematics procedure works Make distinctions between procedural and conceptual knowledge when teaching mathematics concepts and operations to pupils Integrate mathematicsematical ideas from across areas of mathematics
Instructional planning	Accommodate a wide range of abilities in each lesson Create learning experiences that make the central concepts of subject matter meaningful to pupils Create projects that motivate all pupils to participate Deal with learning difficulties so that specific pupil outcomes are accomplished Develop games or puzzles that provide instructional activities at a high interest level Develop instructional materials that build on pupils' experiences, interests and abilities Use pupils' misconceptions to plan instruction
Assessment uses	Give useful and timely feedback to pupils about their learning Help pupils learn how to assess their own learning Use assessment to give effective feedback to parents or guardians Use assessment to give feedback to pupils about their learning Use classroom assessments to guide your decisions about what and how to teach

Table 1.9 (cont’d)

Assessment practice	Analyze and use national or state standards or frameworks for school mathematics Analyze pupil assessment data to learn how to assess more effectively Assess higher-level goals (e.g., problem-solving, critical thinking) Assess low-level objectives (factual knowledge, routine procedures and so forth) Build on pupils’ existing mathematics knowledge and thinking skills

Table 1.10

Estimated effects of OTL in mathematics education courses on MPCK score

	Class participation	Class reading	Solving problems	Instructional practice	Instructional planning	Assessment uses	Assessment practice
Chinese Taipei	3.02*	4.23***	3.72**	3.11*	2.92**	1.23	3.82**
Singapore	4.85	.88	4.09	1.69	-1.95	.72	.65
Spain	0.54	-1.35	-.57	.94	.89	-.74	-1.62
Switzerland	-.26	-1.74*	-.32	-.70	-.41	-.64	-3.76**
United States	-2.96	-2.24	-6.74**	-2.85	-.95	-1.81	.37

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

Table 1.11

Estimated effects of number of topics addressed in general pedagogical courses on MCK and MPCK scores

	MCK		MPCK	
	Gen_ped_top (candidate)	Gen_ped_cou (Institution)	Gen_ped_top (candidate)	Gen_ped_cou (Institution)
Chinese Taipei	-.03	.00	.00	.04*
Singapore	.16**	.17**	.14*	.16*
Spain	.03	.00	.01	-.01
Switzerland	-.01	.03	.04	.02
United States	-.06	.05	-.07	.02

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

Table 1.12

Estimated effects of nature of general pedagogical topics on MCK and MPCK scores

	Social science		Application	
	MCK	MPCK	MCK	MPCK
Chinese Taipei	-4.14	-.09	-.10	1.35
Singapore	30.79**	7.54*	8.13**	6.67*
Spain	7.26*	5.26*	-1.16	-1.39
Switzerland	8.11*	3.19	-5.98	2.61
United States	-.31	-3.41	-6.02*	-5.22*

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

Table 1.13

Lists of items used to construct each OTL variable in general pedagogical courses

OTL	In your teacher preparation program, how frequently did you do any of the following? (Never, Rarely, Occasionally, Often)
Teaching for diversity	Develop specific strategies for teaching students with behavioral and emotional problems Develop specific strategies and curriculum for teaching pupils with learning disabilities Develop specific strategies and curriculum for teaching gifted pupils Develop specific strategies and curriculum for teaching pupils from diverse cultural backgrounds Accommodate the needs of pupils with physical disabilities in your classroom Work with children from poor or disadvantaged backgrounds
Teaching for reflection on practice	Use teaching standards and codes of conduct to reflect on your teaching Develop strategies to reflect upon the effectiveness of your teaching Develop strategies to reflect upon your professional knowledge Develop strategies to identify your learning needs
Teaching for improving practice	Develop and test new teaching practices Set appropriately challenging learning expectations for pupils Learn how to use findings from research to improve knowledge and practice Connect learning across subject areas Study ethical standards and codes of conduct expected of teachers Create methods to enhance pupils' confidence and self-esteem Identify opportunities for changing existing schooling practices Identify appropriate resources needed for teaching

Table 1.14

Lists of items used to construct each OTL variable in student teaching

OTL	During the school experience part of your program, how often were you required to do each of the following? (Never, Rarely, Occasionally, Often)
Connecting classroom learning to practice	Observe models of the teaching strategies you were learning in your <courses> Practice theories for teaching mathematics that you were learning in your <courses> Complete assessment tasks that asked you to show how you were applying ideas you were learning in your <courses> Receive feedback about how well you had implemented teaching strategies you were learning in your <courses> Collect and analyze evidence about pupil learning as a result of your teaching methods Test out findings from educational research about difficulties pupils have in learning in your <courses> Develop strategies to reflect upon your professional knowledge Demonstrate that you could apply the teaching methods you were learning in your <courses>
	To what extent do you agree or disagree with the following statements about the <field experience and/or practicum> you had in your teacher preparation program? (Disagree, slightly disagree, slightly agree, agree)
Supervising teacher reinforcement of university goals for practice	I had a clear understanding of what my school-based <supervising teacher/mentor/instructors> expected of me as a teacher in order to pass the <field experiences/ practicum>. My school-based <supervising teacher/mentor/instructors> valued the ideas and approaches I brought from my <university/college> teacher education program. My school-based <supervising teacher/mentor/instructors> used criteria/standards provided by my <university/college> when reviewing my lessons with me. I learned the same criteria or standards for good teaching in my <courses> and in my <field experiences /practicum>. In my <field experience / practicum> I had to demonstrate to my supervising teacher that I could teach according to the same criteria/standards used in my <university/college> <courses>.
Supervising teacher feedback quality	The feedback I received from my <supervising teacher/mentor/instructors> helped me to improve my understanding of pupils. The feedback I received from my <supervising teacher/mentor/instructors> helped me improve my teaching methods. The feedback I received from my <supervising teacher/mentor/instructors> helped me improve my understanding of the curriculum. The feedback I received from my <supervising teacher/mentor/instructors> helped me improve my knowledge of mathematics content.

Table 1.15

Estimated effects of OTL in student teaching on MPCK score

	Connecting	Reinforcement	feedback	Supervisor_present
Chinese Taipei	.60	-3.30*	.62	3.95*
Singapore	2.44	1.67	4.45*	-7.02*
Spain	-1.13	-1.39	-1.08	4.21*
Switzerland	.25	-.06	-.64	4.27*
United States	-2.39	-.09	-.82	.35

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

Table 1.16
Summary of the statistically significant positive relationships of teacher preparation components to teacher knowledge

		Chinese Taipei		Singapore		Spain		Switzerland		United States	
Hypothesis	Independent variable	MCK	MPCK	MCK	MPCK	MCK	MPCK	MCK	MPCK	MCK	MPCK
1A	number of mathematics content courses	F*	F*	I*	F*/I*	---	---	F*	---	---	---
1B	Geometry	*		---		---		---		---	
	discrete structure & logic	*		---		*		*		---	
	continuity & functions	*		---		---		*		*	
	probability & statistics	*		---		---		*		---	
2A	number of topics addressed in mathematics methods courses	F*	---	---	---	---	---	F*	F*/I*	---	---
	Foundation	---	---	---	---	---	---	*	*	---	---
	Instruction	*	---	*	*	---	*	*	*	---	---
2B	class participation		*		---		---		---		---
	class reading		*		---		---		---		---
	solving problems		*		---		---		---		---
	instructional practice		*		---		---		---		---
	instructional planning		*		---		---		---		---
	assessment uses				---		---		---		---
	assessment practice		*		---		---		---		---
3A	number of topics addressed in general pedagogy courses	---	I*	F*/I*	F*/I*	---	---	---	---	---	---
	social science	---	---	*	*	*	*	*	---	---	---
	Application	---	---	*	*	---	---	---	---	---	---
3B	teaching for diversity		---		---		---		---		---
	teaching for reflection on practice		---		---		---		---		---
	teaching for improving practice		---		---		---		---		---
4A	Student teaching	---	---	---	---	---	---	---	---	---	---

Table 1.16 (cont’d)

		Chinese Taipei		Singapore		Spain		Switzerland		United States	
Hypothesis	Independent variable	MCK	MPCK	MCK	MPCK	MCK	MPCK	MCK	MPCK	MCK	MPCK
4B	connecting classroom learning to practice	▨	---	▨	---	▨	---	▨	---	▨	---
	supervising teacher reinforcement of university goals for practicum,	▨	---	▨	---	▨	---	▨	---	▨	---
	supervising teacher feedback quality	▨	---	▨	*	▨	---	▨	---	▨	---
	Supervising teacher present in classroom	▨	*	▨	---	▨	*	▨	*	▨	---

Note. Statistically significant positive relationships are indicated in the table by *. When the variables are both at the future teacher level and the institution level, the letter F (future teacher) and I (institution) indicate the level at which the positive relationship exists. Potential relationships that are not significantly positive are indicated by ---. Gray spaces indicate that the implied relationship was not modeled.

Table 1.17

Standard deviations of MCK and MPCK and the number of significantly positive relationships between teacher preparation components and knowledge for five countries

Country	MCK standard deviation	MPCK standard deviation	The number of relationships
Chinese Taipei	84.23	68.39	16
Singapore	72.39	73.32	14
Spain	56.56	63.24	13
Switzerland	65.00	62.06	5
United States	70.01	67.60	1

Table 1.18

Mean scores of MCK and MPCK and the number of significantly positive relationships between teacher preparation components and knowledge for five countries

Country	Mean score of MCK	Mean score of MPCK	The number of relationships
Chinese Taipei	623	592	16
Singapore	586	588	14
Switzerland	548	539	13
Spain	481	492	5
United States	518	544	1

Chapter 2. Using response time to detect item pre-knowledge in computer-based testing

Item pre-knowledge means an examinee knows some items in item pool and their answers before taking the test. Such a situation is possible for computer-based licensure examinations for at least three reasons. First, substantial stakes are associated with licensure examinations, as the passing of such a test usually makes the examinee eligible to work in a given field, or qualifies the individual for a higher paid position (Smith & Davis-Becker, 2011). As a result, examinees might be motivated to gain access to some of the items before they take the test. For example, the examinee may ask a colleague who has taken the exam earlier about items they remember, search the internet for stolen content, or explore the offerings of training schools which provide the content as part of the test-preparation materials (Smith & Davis-Becker, 2011). Second, the problem of item exposure is further exacerbated because items in computer-based testing usually remain operational for some time in order to return the investment on item development (van der Linden & van Krimpen-Stoop, 2003). Examinees may try to use this time to memorize and share items with others. Third, many licensure examinations are delivered on-demand and have large volumes, making them more vulnerable to item exposure problems (Cohen & Wollack, 2006). Item pre-knowledge can be a threat to the validity of the inferences from examinations because it is unclear whether the examinee passed the test because of the competency he/she has in the given field or because he/she knows some items before taking the test. Therefore, it is important for testing programs to identify potentially compromised items by monitoring examinee response behavior (Smith & Davis-Becker, 2011).

A traditional way to detect item pre-knowledge is to use response data to do person-misfit analysis (McLeod & Lewis, 1999; Veerkamp, 1996; van Krimpen-Stoop & Meijer, 2000; Meijer

& Sijtsma, 1995). If a low-ability examinee answered more difficult questions correctly than would be expected by chance, it may indicate that this examinee knew these items before he/she took the test. However, this method usually has a low detection rate and a high false alarm rate and cannot be used in operational testing. One possible way to counter these problems is to complement analysis of response data with analyses of response time data.

In computer-based testing, each examinee's response time on each item is automatically recorded. More importantly, item pre-knowledge may be detected by unexpected response times. That is, examinees who know some of the items prior to the test may answer them more quickly than is typically the case. The purpose of the research reported in this essay was to detect examinees' item pre-knowledge behavior and detect compromised items in an item pool using response and response time data for a computer-based licensure test. Specifically, there are three research questions addressed in this essay:

- To what extent do examinees have item pre-knowledge in a large-scale licensure exam?
- To what extent are there potentially compromised items in this exam item pool?
- What is the power of the methodology to detect item pre-knowledge?

In the first section of this essay, I review the research literature on detecting item pre-knowledge. The second section describes the models I used in the study. The third section introduces my research methods, including samples and procedures. In the fourth section, I present model validation in terms of whether the data fulfill the assumptions under the models. The fifth section presents the results for item pre-knowledge and compromised items detection. In the last section, I describe the results for a simulation study and provide conclusions.

Literature review

In this section, I review research on detecting item pre-knowledge. Several researchers have tried to use response times to detect possible aberrances in examinee behavior. van der Linden and van Krimpen-Stoop (2003) used response and response times to detect two different types of aberrances: item pre-knowledge (unexpectedly correct responses and unexpectedly short response times) and speededness (unexpected incorrect responses and unexpectedly short response times at the end of the test due to running out of time). They used classical procedures and Bayesian posterior predictive checks in simulation studies. For classical checks, the detection rate is .30 and the false-alarm rate is .05. For Bayesian checks, the detection rate doubled relative to the classical checks, but at the cost of a considerable increase in the false-alarm rate.

Meijer and Sotaridona (2006) used effective response time to detect item pre-knowledge. Effective response time is defined as the time required for an individual examinee to answer an item correctly. To investigate the power of the methods, a random sample of examinees from the actual data set was selected and their response time was changed to one-half or one-fourth of the original response time on one-half or three-fourths of all the items they responded to. The method is sensitive to the amount of time reduced due to item pre-knowledge. For example, the method has high power to detect item pre-knowledge for examinees who know half of the items and whose quick response is equal to one-fourth of the normal time (the detection rate is .944). However, it may be not realistic that examinees can have access to half of the items on the test.

van der Linden and Guo (2008) used a hierarchical framework to detect two aberrant response-time patterns: item pre-knowledge (unexpectedly short response times) and taking tests only for the purpose of memorizing the items (unexpectedly long response times). The procedure was illustrated using a data set for the Graduate Management Admission Test (GMAT). The procedure identified 1.69% of examinees who spent less time than expected and 2.25% of examinees who spent more time than expected. These percentages are close to the nominal significance level of the test, which means that the test takers generally behaved quite regularly according to the response time model and that cheating or item compromise was certainly not a structural problem for the GMAT. To make sure that the results were not due to a lack of power of the method, a power study was conducted using simulated cheating behavior. The response time for known items was set to 10, 20, or 30 seconds. The detection rate for 10 seconds was quite high (0.8); for 20 seconds, it was acceptable (0.4); for 30 seconds, detection rate was low (0.2). It should be noted, however, that these rates were only for a single item. If a test taker knows more than one item, the power of the procedure would go up immediately. Since this method has a good detection rate. My study used the models from this study.

Models used

I used a hierarchical framework proposed by van der Linden (2007) for modeling response and response times[6].

First level models: two separate models for the response and response times

Response model: two-parameter logistic model

$$P(u_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{e^{1.7a_i(\theta_j - b_i)}}{1 + e^{1.7a_i(\theta_j - b_i)}}$$

where $\theta_j \in \Re$ is the ability parameter for the test taker j, $b_i \in\triangleleft \Re$ is the difficulty parameters for item *i*, $a_i \in \Re^+$ is the discrimination parameter for item *i*.

[6] Essay two will only use the first level model. Essay three will use both the first level and second level models.

Response time model: a lognormal model (van der Linden, 2006)

Suppose test taker *j* operates on item *i* at speed τ_j ($\tau_j \in \Re$). We take the response time to be a random variable T_{ij}. The times t_{ij} observed during the test are realizations of these random variables.

$$f(t_{ij};\tau_j,\alpha_i,\beta_i)=\frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\exp\{-\frac{1}{2}[\alpha_i(\ln t_{ij}-(\beta_i-\tau_j))]^2\}$$

where $\beta_i \in \Re$ represents the time intensity of item i; a larger value means systematically more time spent on the item by the test takers. $\alpha_i \in \Re^+$ represents the discriminating power of item i, the reciprocal of the standard deviation of the response-time distribution; a larger value for α_i means less dispersion of the log times on item i for all test takers. The residuals for this model are the following:

$$\ln T_{ij} = \beta_i - \tau_j + e_i,\ e_i \sim N(0,\ \alpha_i^{-2}).$$

$$e_i = \ln t_{ij} - (\beta_i - \tau_j),\ e_i \sim N(0,\ \alpha_i^{-2}).$$

$$e_i = \alpha_i\ (\ln t_{ij} - (\beta_i - \tau_j)),\ e_i \sim N(0,1).$$

Second level models

The previous two models are for the marginal distributions of the response and response time by a test taker on an item. Although these distributions are assumed to be independent, response and response times may show substantial correlation across test takers and/or items. To allow for such dependencies, two separate second-level models are required, one for the joint distribution of the person parameters, and another for the joint distribution of the item parameters.

Population model:

$$(\theta, \tau) \sim MVN(\mu_{\theta\tau}\, \Sigma_{\theta\tau})$$

with $\mu_{\theta\tau}$ the vector of the population means of θ and τ and $\Sigma_{\theta\tau}$ their covariance matrix.

Item domain model:

$$(b, \alpha, \beta) \sim MVN(\mu_{b\alpha\beta}, \Sigma_{b\alpha\beta})$$

with $\mu_{b\alpha\beta}$ the vector of the means of all item parameters and $\Sigma_{b\alpha\beta}$ their covariance matrix.

The full two-level model is represented in Figure 2.1.

For model identification, the following constraints are suggested:

$$\mu_{\theta} = 0, \sigma_{\theta}^{2} = 1, \mu_{\tau} = 0$$

If the time is measured in seconds, the scale of $\ln t_{ij}$ is fixed and so is the scale of τ or any other of the time parameters.

Parameters are estimated using a Bayesian approach with Gibbs sampler as a Markov chain Monte Carlo (MCMC) method for sampling from the posterior distribution of the parameters (van der Linden, 2008). A software package written in R, Cirt, is available for parameter estimation (Fox, Entink, and van der Linden, 2007). The input files are N×K matrix of the responses (N: the number of persons; K: the number of items) and N×K matrix of the log response times. The output files include all model parameters and some of the model fit criteria.

Method

Sample

To detect item pre-knowledge, two samples are needed. The first sample, including N×K matrix of the responses and N×K matrix of the log response times, is used for item calibration. This sample is from the early stage of the operational test, where there are no compromised items. The second sample is from the later stage of the operational test that is subject to content exposure problem. For this sample, I only need to estimate person parameters since all item parameters were assumed to be known from previously calibrated items.

The data used in this study are from a large-scale computer-based licensure examination. Two non-consecutive years of data were used – the first and third year of exam administration. For the early sample, 992 candidates taking a 185 multiple-choice item exam in the first six months of administration (January-June of 2010) were used to estimate the model. It was assumed that, as those candidates have taken a completely new exam, there was no compromised content, and the recovered parameters are the “true” item parameters. These parameters were then applied to the second sample of 1,172 candidates taking the test in early 2012 to detect possible item pre-knowledge. There were 111 items in common between the two samples. Slightly over 10,000 candidates took the test between June 2010 and December 2011.

Procedures

The unexpected short response time is detected by estimated residual log response time:

$$a(\ln t_{ij} - (\beta_i - \hat{\tau}_j)), i = 1, ..., K; \ j - 1, ..., N$$

These residuals have an approximate standard normal distribution. Because the speed parameters are estimated for the full test, increase in the actual speed on subsets of items manifests itself by larger negative values for the residuals. A response time to an item was flagged as aberrant when its residual was smaller than -1.96. Even though there are a lot of related significant tests and should use Bonferroni correction (using much smaller critical value, which will indicate a smaller number of flagged response times), I used -1.96 to be more conservative and wanted to identify more flagged response times. Then I can check how often the response times for the same person were flagged to detect suspect examinees. I can also check how often the response times on the same items were flagged to detect compromised items.

Model validation

There are several assumptions under the models I will use. Before I can use the models to predict the reasonable response times, then compare them to the observed ones and identify unexpected ones, I have to test whether the data fit the model first. The main assumptions under the models described in the previous section include the following.

1. The responses fit the two-parameter logistic (2PL) model

I used IRTPRO to check whether the response data fit the 2PL model according to item level diagnostic statistics. The results showed that there are only two items among 111 items showing misfit, which indicates that most of the responses fit the 2PL model.

2. The response times fit the lognormal distribution

Figure 2.2 shows a histogram of the response times for item 72015. The x-axis presents the response times in seconds while the y-axis is the number of examinees. As is typical of response-time distribution, the data are unimodal and positively skewed. In order to visually show whether the response times fit the lognormal distribution, I plotted the observed cumulative density function (CDF) and predicted CDF for lognormal distribution for each item. The parameters of the lognormal density were estimated by taking the mean and standard deviation of the natural logarithm of 2010 response time data. If the observed CDF and predicted CDF are close to each other, the response time data fit lognormal distribution. Figure 2.3 and Figure 2.4 showed the plots for two items. Results can also be shown in the form of double probability plots (Schnipke & Scrams, 1999): for each unique observed response time, I plotted its observed cumulative probability against its predicted cumulative probability. If the data are fit well by a lognormal distribution function, the points will be close to 45^{o} diagonal (perfect fit). Figure 2.5 and Figure 2.6 showed double probability plots for the same two items shown in Figure 2.3 and Figure 2.4. From Figure 2.5 and Figure 2.6, there are minor deviations from the diagonal. Although Figure 2.3, 2.4, 2.5 and 2.6 give us the visual presentation of misfit, we cannot determine whether such deviation is trivial or should be of concern. In order to quantify misfit, I calculate the root mean squared error (RMSE) for each item based on a lognormal distribution using 2010 data as the exploratory sample. RMSE is based on the square root of the mean squared difference between the observed and predicated CDF at every 5th percentile in the observed CDF from the 5th to the 95th. Large values of RMSE indicate poor fit. Table 1 shows RMSE for the lognormal distribution for each item.

From Table 2.1, most of the items have RMSE smaller than 0.05. To test whether response time data fit other distributions better, response times for the 2010 sample were fit with three distribution functions (lognormal, gamma, and normal). To show how well each distribution fit on each item, the values of RMSE for each distribution are shown for each of the 110 items in Figure 2.7.

As shown in Figure 2.7, the lognormal distribution provided the best fit on most of the items and the normal distribution provided the worst fit. The gamma distribution was in the middle in terms of how well it fits each item. Table 2.2 shows the mean RMSE for each of the three distributions across items, as well as the minimum (best) and maximum (worst) values of RMSE. As shown in Table 2.2, the lognormal distribution provides the best fit, followed by the gamma. The normal distribution provides the worst fit overall.

In the second set of analyses, response times for the 2012 sample (the confirmatory sample) were fit with the three distribution functions using the parameter estimated obtained from the

2010 sample. To summarize the fit for each of the three distributions for 2012 sample, I calculated RMSE, as I had done for the 2010 sample. Table 2.3 shows the mean RMSE for each of the three distributions, as well as the minimum (best) and maximum (worst) values of RMSE. As in the exploratory sample, the lognormal distribution provides the best fit overall.

I also used a Chi-square goodness-of-fit test to test the null hypothesis that the log response times for each of the 111 items are a random sample from a normal distribution with mean and variance estimated from the log response time for each item. Among 111 items, the null hypothesis cannot be rejected in 106 items after Bonferroni correction. It means that most items fulfill this assumption.

Results

Descriptive data analysis

Figure 8 shows the distributions of response times in seconds for one item in 2010 and 2012. While the general pattern follows the typical response time distribution, there is a little bump at the left for each distribution. These are the examinees who responded to the item very quickly. For 2010, about 40 examinees responded to this time in less than 40 seconds while the median time was 400 seconds. It is possible that some examinees did not want to spend a lot of time on one item and quickly guessed an answer and went to the next item. Therefore, almost half of them got it wrong. In 2012, there are even more examinees responding to this item quickly. However, a close look reveals that the increase is greatest in the blue part, which are incorrect answers. Therefore, even though there are more examinees answering the item quickly in 2012 than in 2010, it is not due to item pre-knowledge because if an examinee knows an item before the test, he/she would not only answer it quickly, but also correctly. While Figure 2.8 does not reflect item pre-knowledge, Figure 2.9 tells us another story.

In Figure 2.9, the number of examinees in the first bar increased in 2012 and most of them answered the item correctly. It may indicate that some examinees knew this item before they took the test after the item had been exposed for two years. Figure 2.9 visually shows a possible compromised item. The statistical procedure used in this study will identify such items more efficiently and accurately than plotting.

The procedure used in this study can not only detect compromised items, but also identify examinees who answer some questions unexpectedly quickly relative to his/her own speed. For an examinee who is regularly fast, the response time needs to be extremely short in order to be detected as unexpectedly quick while for an examinee who is extremely slow, a typical response time may seem quick for this examinee. Figure 2.10 and Figure 2.11 show the fastest examinee's and slowest examinee's response times for 110 items against the median response times.

For the fastest examinee in Figure 2.10, his/her response times for almost all items are shorter than median response times (the blue line is below the red line). At the same time, this examinee also follows the pattern of response times for each item, which means that he/she spent a little bit more time on the items that others spend a lot of time on. For the slowest examinee in Figure 11, his/her response times for a lot of items are longer than median response times (the blue line is above the red line). Figure 2.12 and Figure 2.13 show the residuals for the same two examinees.

In Figure 2.12, the *x*-axis is 110 items, and the *y*-axis is residuals. A residual above 0 means the examinee spent more time on this item than expected based on the person's speed and the item time intensity. A residual below 0 indicates the examinee spent less time on this item than predicted from the model. The green line indicates the correct answer while the red line indicates the wrong answer. For the fastest examinee, it is not surprising that there are a lot of negative residuals. However, among those negative residuals, there are none less than -1.96. Figure 12 shows that this examinee is regularly fast and not extremely fast on some items. Therefore, there is no indication for item pre-knowledge. In Figure 2.13, for the slowest examinee, there are a lot of positive residuals and some negative residuals. And for negative residuals, there are none as extreme as -1.96. Figure 2.14 shows a case that has a lot of large negative residuals.

As shown in Figure 2.14, this examinee spent more time than expected on the first 65 items. Then suddenly he/she changed the pattern and responded very quickly on the rest of the items. A lot of residuals are smaller than -1.96. However, among them most are incorrect. The proportion of correct responses is about the same as would be expected by chance. Therefore, even though there are a lot of large negative residuals, they are not an indication for item pre-knowledge since most of them are wrong. It is possible that the examinee thought he/she may run out of the time (actually he/she only spent 200 minutes on the test while the time limit for the test is 300 minutes). Another possibility is that this is a person coming from test preparation school and he or she only wanted to get information about the test. After he/she obtained enough information, he/she just did random guessing and tried to finish the test quickly. From Figure 2.14, we can see that response times can be used to detect various kinds of aberrant patterns, such as speededness or rapid guessing. However, the focus of this study is to detect item pre-knowledge and compromised items. The next section presents the results for detecting item pre-knowledge.

Results for detecting item pre-knowledge and compromised items

I checked residual log-response time for all examinees in the 2010 and 2012 samples. The reason I checked residual log-response time for examinees in the 2010 sample is that the number of flagged response times in 2010 reflects type 1 error because in 2010 there should be no item pre-knowledge cases. Then I compared the number of flagged response times for examinees in the 2010 and 2012 samples. Using flagged response times in 2010 as a baseline, I controlled for the type 1 error. The results showed two examinees in 2012 who had significantly higher numbers of flagged items than the 2010 baseline. Figure 2.15 shows the residuals for one of the examinees.

Note large negative residuals indicate faster responses than typical for the examinee. This examinee spent an extremely short time on five items (items 3, 9, 24, 52, and 80) relative to the time spent on the other items. For example, the residual log-response time for item 24 was -4.1. As the model assumed a standard normal distribution of the residuals, the probability of such negative residual by chance alone is 0.000021. The estimated time intensity of this item was β = 4.63, which means that the median time for a test taker of average speed (τ = 0) taken on the item was 102 seconds on the regular time scale. This examinee spent 3 seconds on this item even though the individual's estimated speed is lower than average overall (τ = -0.14). The examinee responded to all five items correctly and only responded correctly to 60% of the remaining 106 items. As these items are spread throughout the test, the pattern is consistent with possible pre-knowledge of these five items rather than speededness or loss of motivation. Another examinee spent an extremely short time on three items and all of them are correct.

To evaluate whether there were compromised items in the 2012 sample, I compared the number of flagged response times for each item for the 2010 and 2012 samples using a paired *t*-test. The result in Table 2.4 showed that there were significantly more flagged response times in the 2012 items than in the 2010 items (*t*=2.943, p=.004).

After looking at every item in 2010 and 2012, two items showed a significant increase in the number of flagged residuals. Figure 2.16 shows one of the two potentially compromised items I discovered. For this item, there are 34 flagged response times, 26 of which (76.5%) are correct. According to Wise and Kong (2005), the accuracy of the rapid-guessing responses should not exceed chance level, as this clearly does. The results indicate that the item may be compromised. An additional examination of the item's content reveals a long stem with a memorable story, which might add to the ease of knowledge transfer between examinees.

In conclusion, after comparing with the baseline in the first year, this study found item pre-knowledge in the third year to be minimal, with two items (out of 111) potentially exposed, and two candidates (out of 1,172) showing some indication of pre-knowledge on multiple items. To make sure that the results were not due to a lack of power of the method, a study was conducted using simulated data (next section). The item pool in this study was the same pool for the large-scale licensure test. And item pre-knowledge that was simulated was suggested by my knowledge of the operational features of the test.

Results for power study

The purpose of power study is to generate item pre-knowledge (short response time and correct response) data and determine whether the estimation procedure can accurately identify these cases. Since in simulation study, we know how many "true" items were characterized by pre-knowledge in the data, we can calculate the detection rate and see the power of the procedure.

Conditions for simulation: (1) the number of items with pre-knowledge in a 100-item test: 1, 5, 10, 30, 50, and 100; (2) the number of examinees with pre-knowledge among 1,000 test-takers: 10, 50, and 100. For each combination of (1) and (2), the number of replications was equal to 10. The reason for various conditions is to see whether the power of the procedure is sensitive to these conditions.

Procedures for simulation: (1) generate the regular response times and responses (1,000 examinees by 100 items). The response times on the regular items were drawn from the following response time model based on the parameters estimated from 2010 sample. The responses were drawn from a two-parameter logistic model based on the parameters estimated from the empirical sample.

(2) Change some regular response times to extremely short response times according to the condition. For example, for the condition (10 examinees know 5 items), I changed 10 examinees' regular response times on 5 items to very short response times (each examinee may know a different set of 5 items). Then I marked the locations for the change. However, what is a reasonable time for extremely short response times? For the item examinees already know, the response times should be short, but at the same time they should be long enough for them to recognize the item, check whether it has been changed in some minor way, and enter the response. In order to generate realistic time ranges for item pre-knowledge, I used the response time distribution of one compromised item I identified in 2012 data. Figure 2.17 shows the histogram of this item's response times.

As indicated in Figure 2.17, 32 examinees in the first three bars (5, 10 and 15 seconds) may have item pre-knowledge. The mean and standard deviation for these 32 examinees' log response times is 1.5636 and 0.50449. Therefore, log response times for compromised items are generated from a normal distribution with mean and standard deviation from these 32 examinees.

(3) Change the responses to correct. I changed all responses on these same locations that response times changed to be correct whether they were correct or incorrect before.

(4) I used the procedure to identify these item pre-knowledge cases and flagged the cases where the residual was smaller than -1.96. Then the number of flagged cases is divided by the number of “true” cases (For example, for the condition that 10 examinees know 5 items, the total number of “true” cases is 50).

The results for the simulation study are in Table 5. Since the procedure is not sensitive to the number of examinees knowing the item, after 10 items, the simulation study only focused on 100 examinees (10 or 50 examinees knowing 30, 50, and 100 items were marked as N/A in Table 2.5).

From Table 2.5, it is clear that when examinees know 10% of the items or less, the detection rate is about 2/3, which means that if there are 100 “true” item pre-knowledge cases, the procedure can detect 67 cases. This detection rate is promising in the field of psychometrics. When examinees know 30% of the items, the detection rate declines to less than 0.5. When examinees know 100% of the items, the detection rate is close to the level of significance, indicating no power at all. This result makes a lot of sense because when examinees know all of the items, he/she will respond to all items very quickly and his/her speed estimate will be very high. The residuals will not be very negative due to the extremely high speed. Since it is not very realistic for examinees to know more than 30% of items before taking a given test, the method shows promise in identifying potentially exposed items as well as candidates who may have gained pre-knowledge of items.

Conclusions and implications

Checking the response behavior of test-takers for possible aberrances is one of the prime methods of quality control in the testing industry (van der Linden & Guo, 2008). As argued in this paper, item pre-knowledge can be one thread in the informantion related to the validity of the inferences from test results. This study used a response time model proposed by van der Linden (2007) to detect item pre-knowledge and compromised items in the item pool for an operational licensure examination. The results showed two items (out of 111) potentially exposed, and two candidates (out of 1,172) showing some indication of pre-knowledge on multiple items. Since the power of the procedure is about 2/3, we should expect there may be one or more compromised items and one or more candidates having item pre-knowledge that are false negatives.

This study has different implications for a) examinees with possible item pre-knowledge and b) compromised items. For examinees with possible item pre-knowledge, we have to use the results with caution. Even though the simulation study showed quite satisfactory power for the proposed checks, there are other explanations for aberrant response times besides item pre-knowledge, and blind conclusions from statistical significant log response time residuals could easily be wrong (van der Linden & Guo, 2008). After detecting examinees with possible item pre-knowledge, careful qualitative analyses are needed, such as a review of the reported irregularities during the testing session or a look at the video of the test takers when they were taking the test. The evidence from psychometric analysis alone is not strong enough to invalidate the test takers' test scores.

For compromised items detected, we can take a more conservative attitude. The simplest response is to delete these items from the item pool and never use them in an operational test again. It is possible that there are some false alarms, but it is wise to delete an item rather than taking a risk. Furthermore, the implications from this study can help prevent item pre-knowledge at the item development stage, too. For example, after detecting the compromised items, we can examine the content of these items and identify characteristics of them that are more easily exposed to the test takers' population.

For two items detected in this study, I have looked at the actual content of these two items and found out that they are very memorable. For one item, there was a long story in the stem that is easy to memorize and communicate. A colleague may tell another colleague, "If you see the Uncle Joseph story, the answer is B." For another item, there was an uncommon phrase put in quotation marks. Therefore, item development guidelines can be devised based on these characteristics and prevent possible item pre-knowledge in item development. This does not mean that Uncle Joseph cannot be used in the item any more. It simply means that the item developer should keep this issue in mind when they write items and if they do refer to Uncle Joseph, they should consider using it in more than one item.

APPENDICES

Appendix 2.1

Figure 2.1

A hierarchical framework for modeling responses and response times.

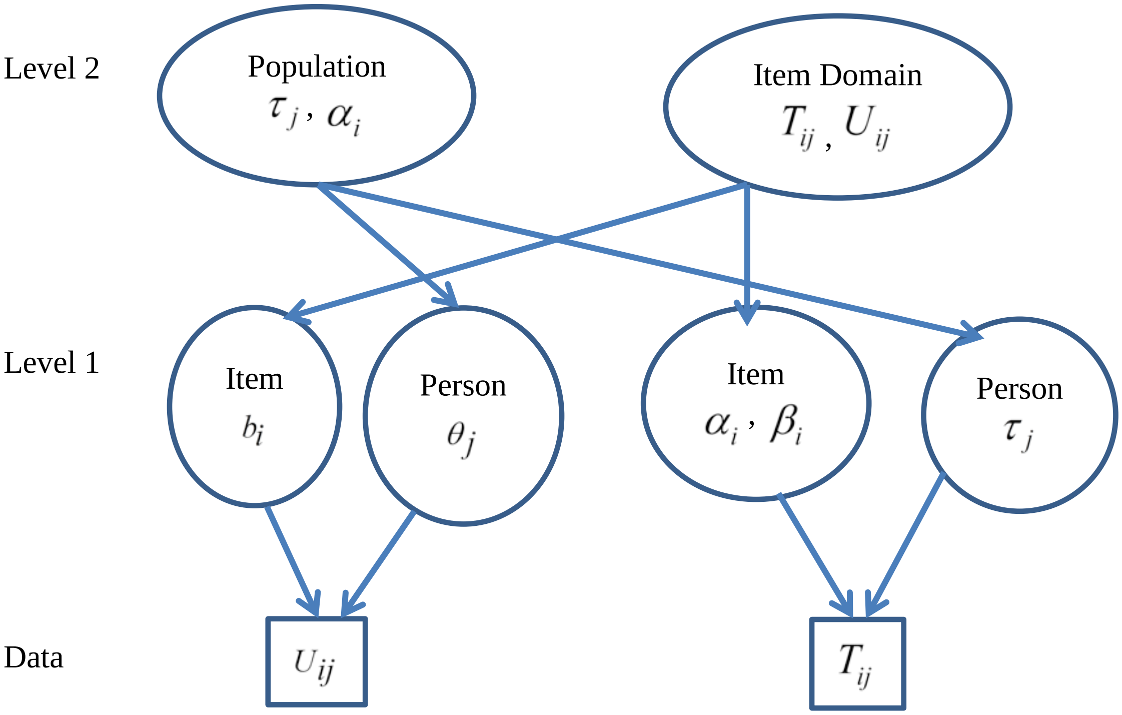


Figure 2.2

A typical histogram of response times

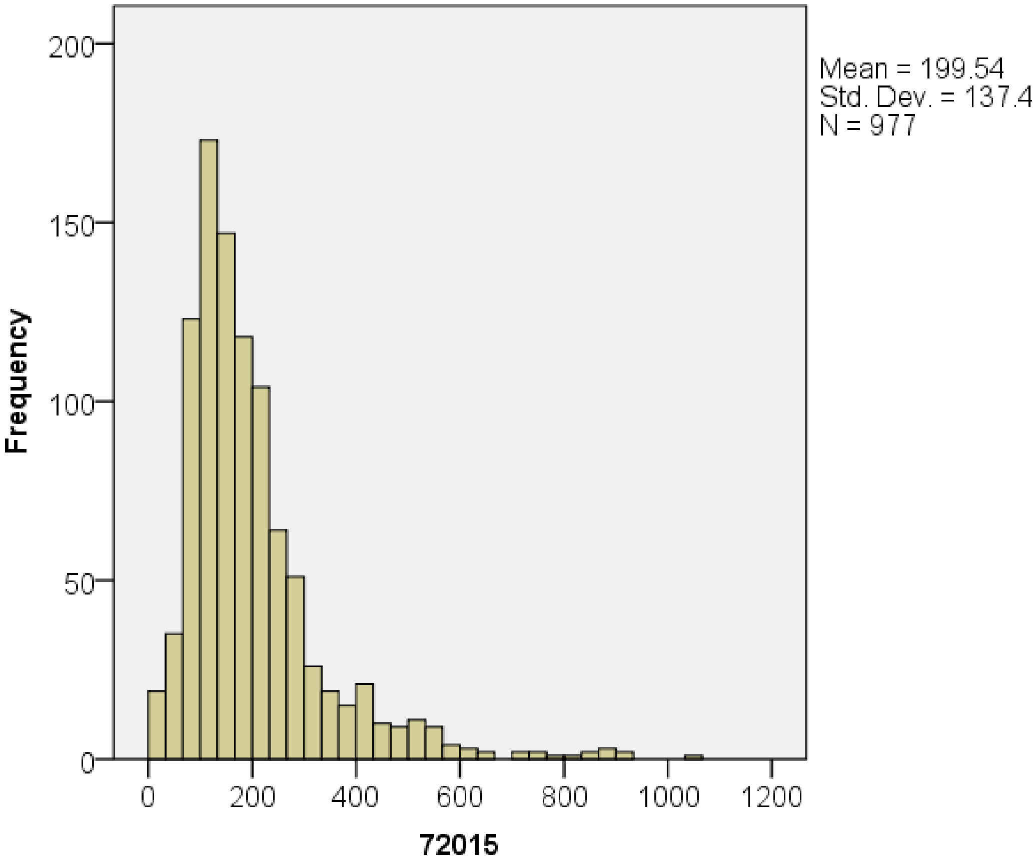


Figure 2.3

The observed CDF and predicted CDF for item 1

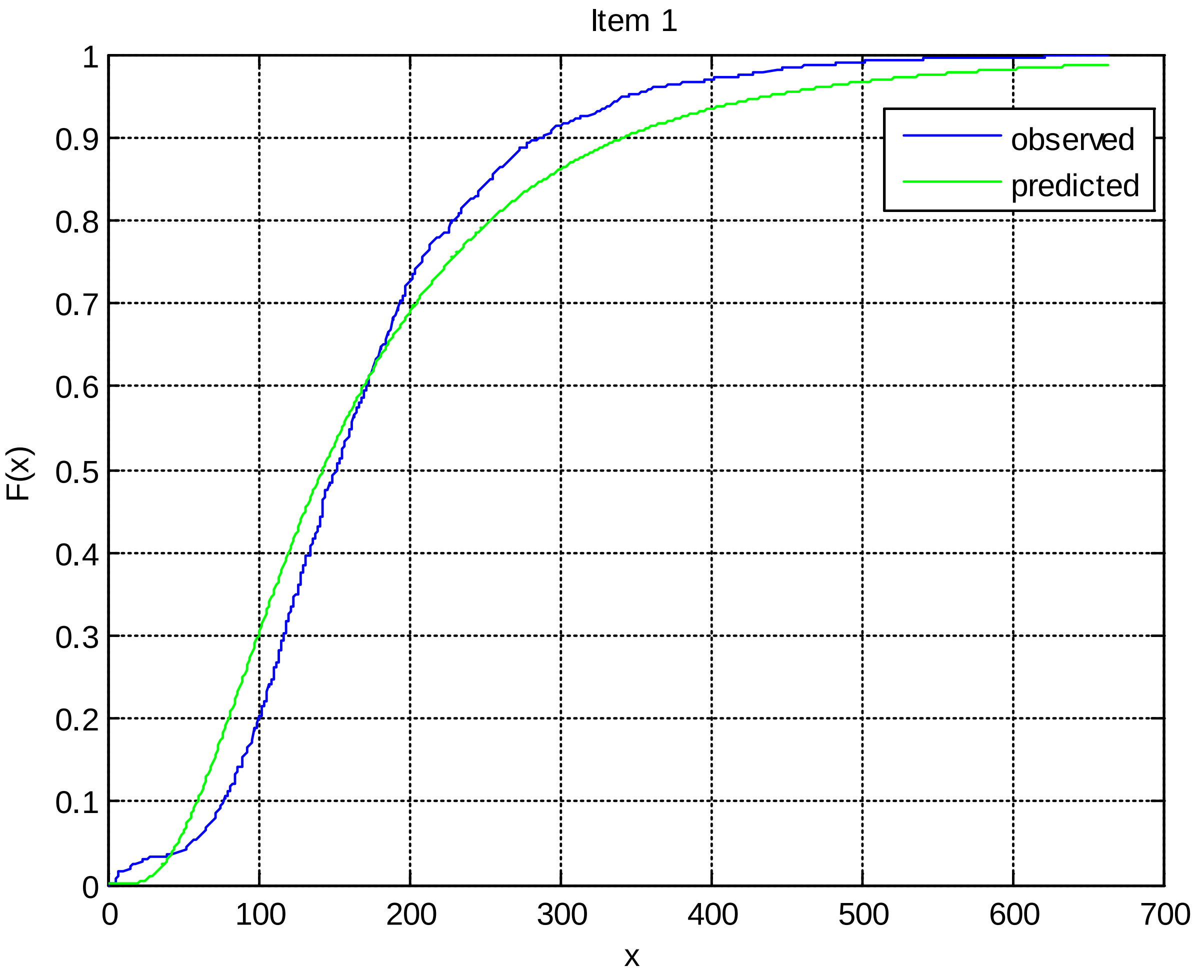


Figure 2.4

The observed CDF and predicted CDF for item 10

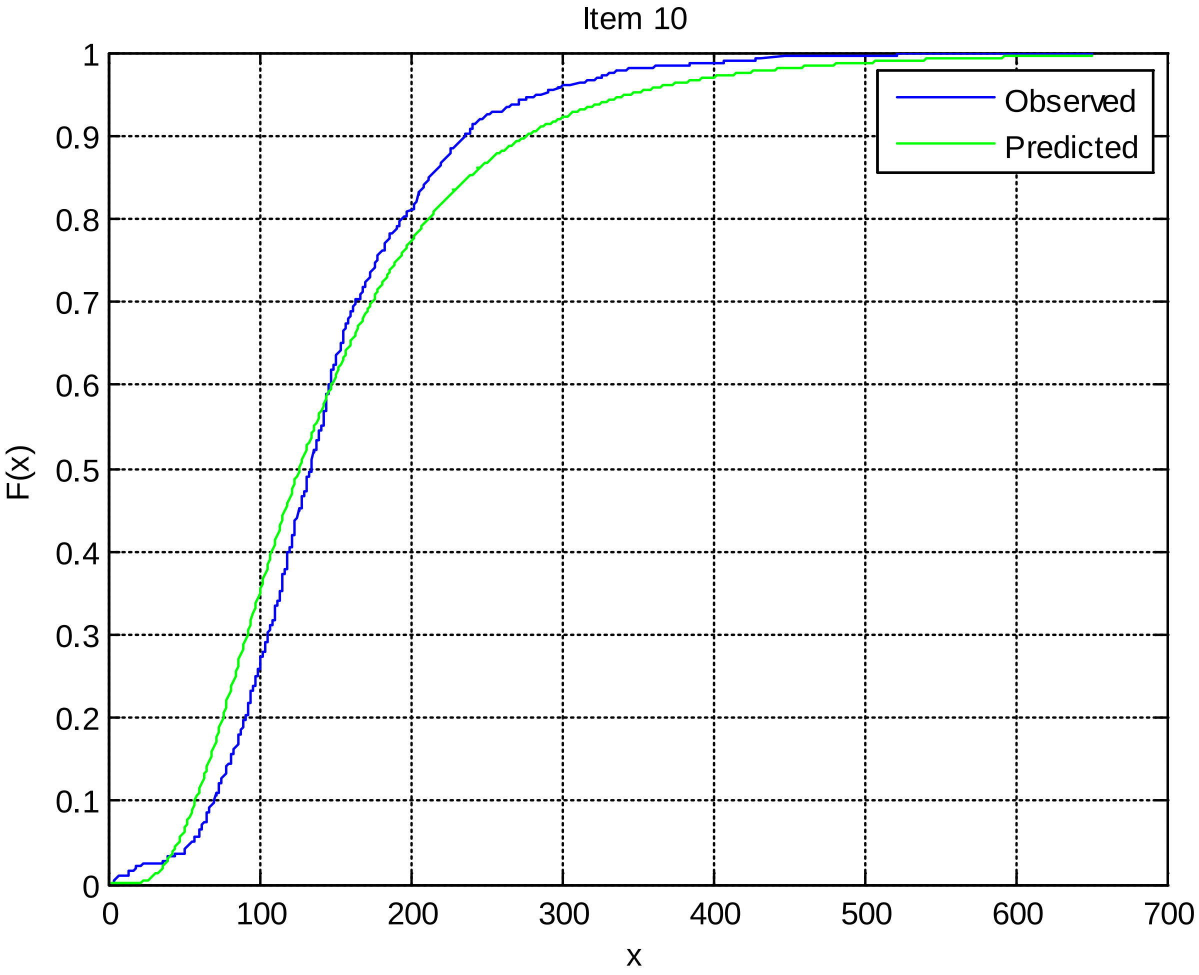


Figure 2.5

Double probability plot for item 1

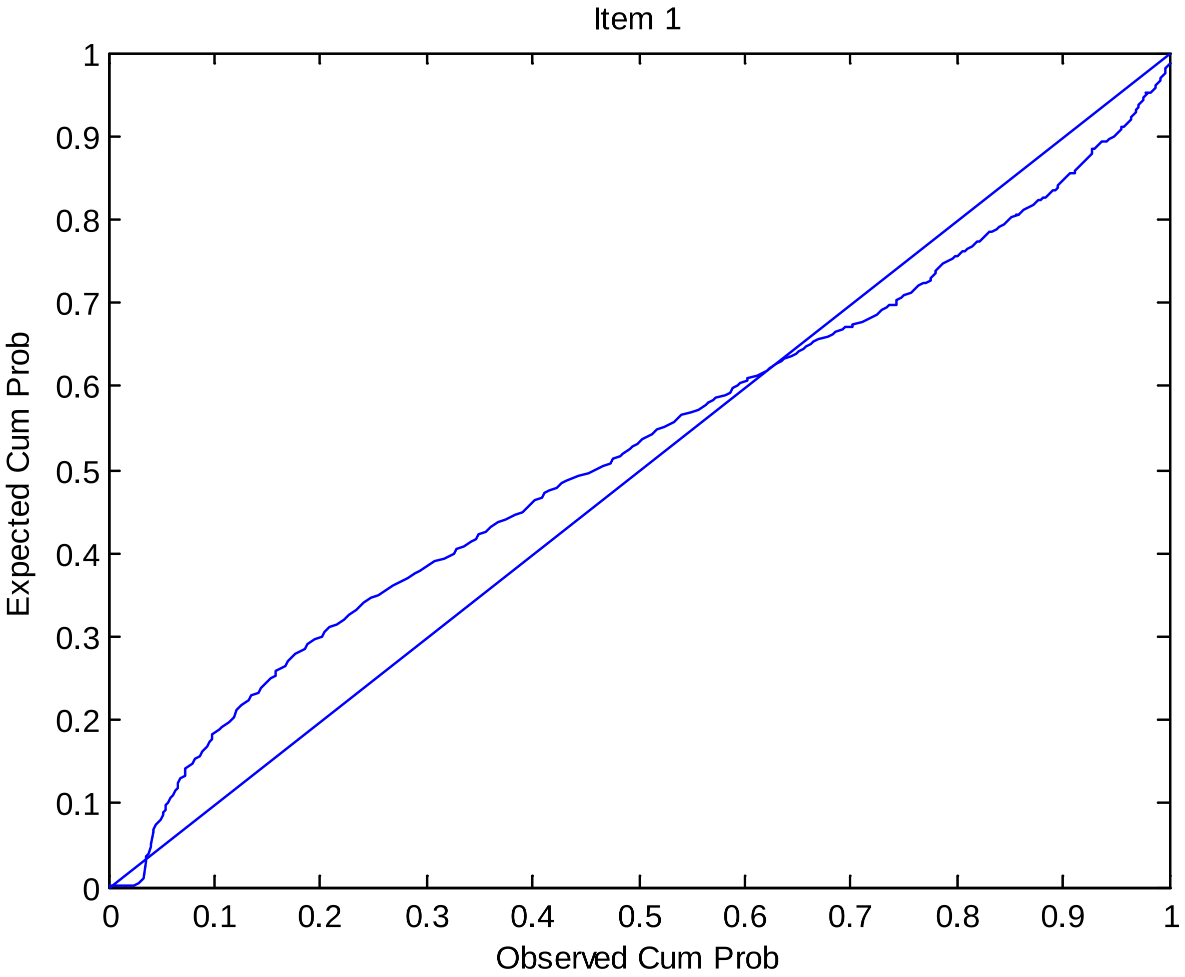


Figure 2.6

Double probability plot for item 10

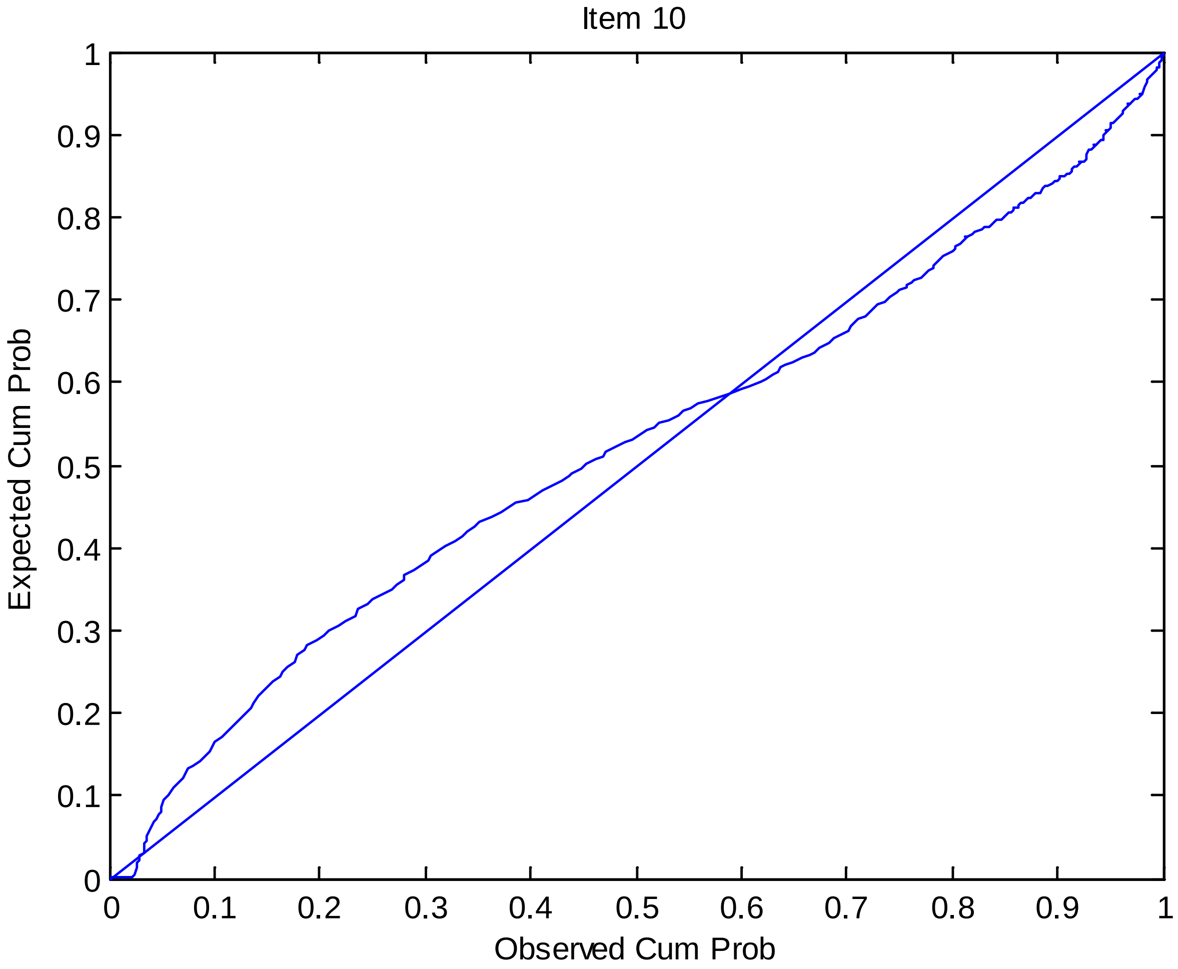


Figure 2.7
RMSE for three distributions for 2010 sample

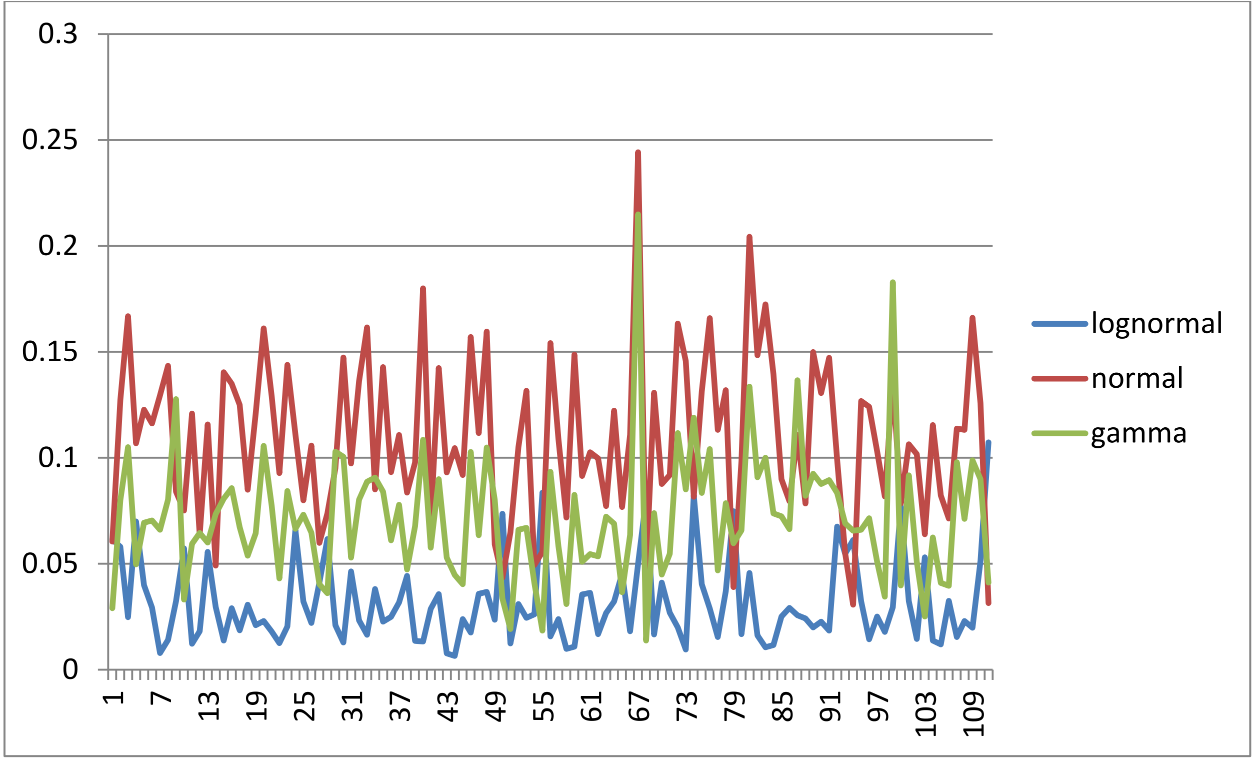


Figure 2.8

The distributions of response times in seconds for item 89121 in 2010 and 2012

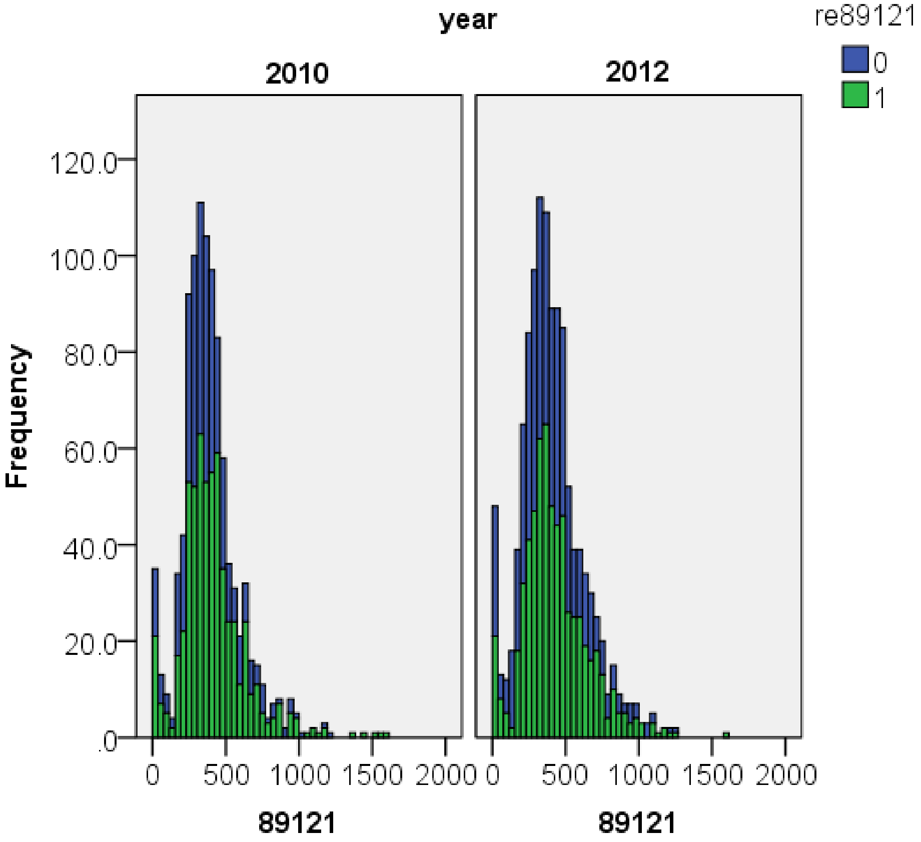


Figure 2.9

The distributions of response times in seconds for item 71965 in 2010 and 2012

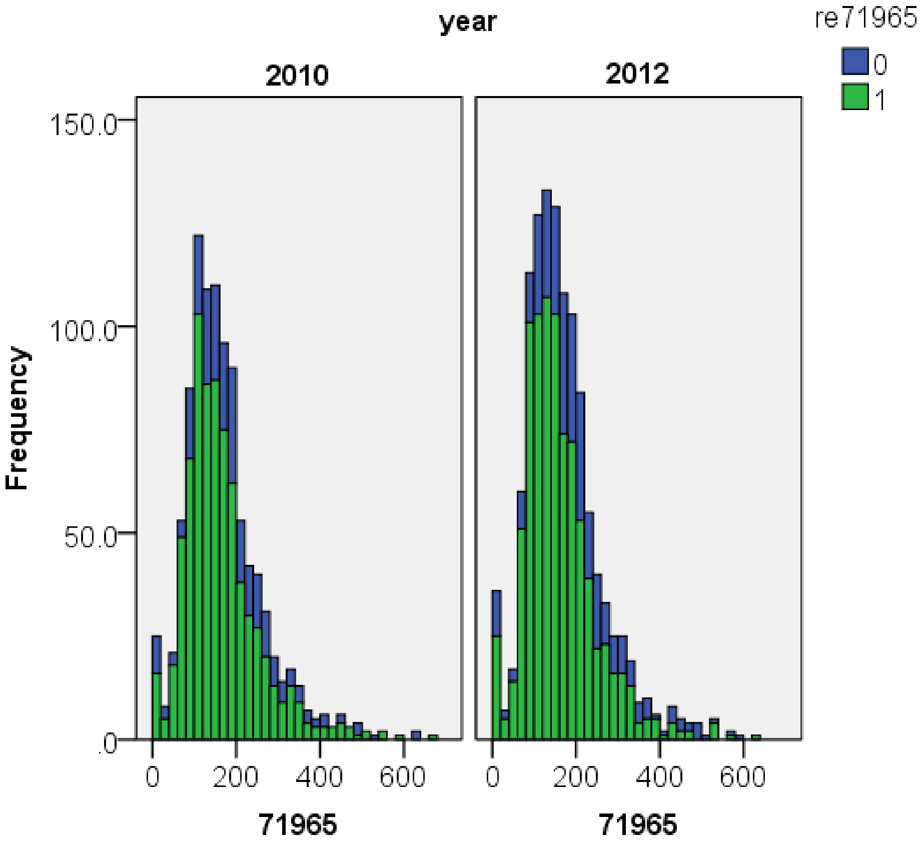


Figure 2.10

The fastest person's response times against median response times

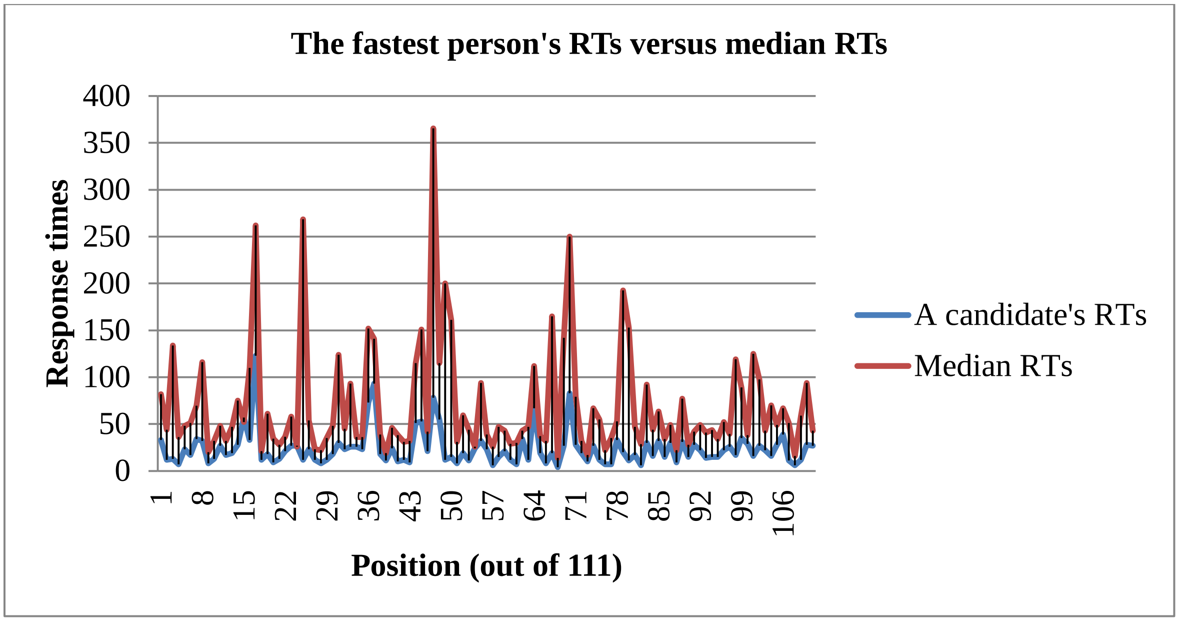


Figure 2.11

The slowest person's response times against median response times

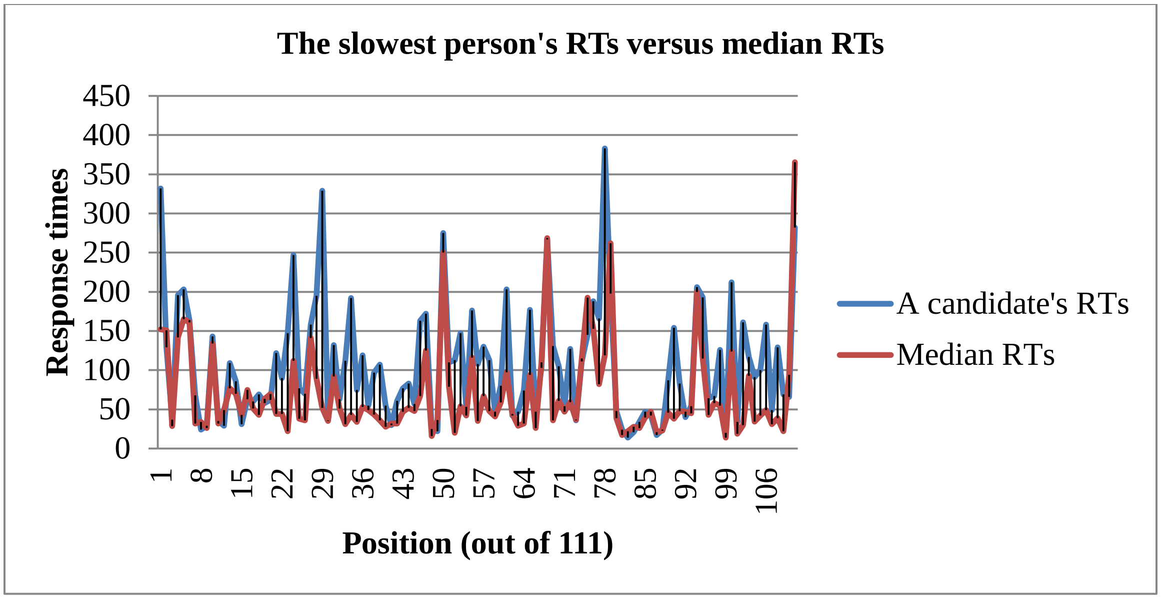
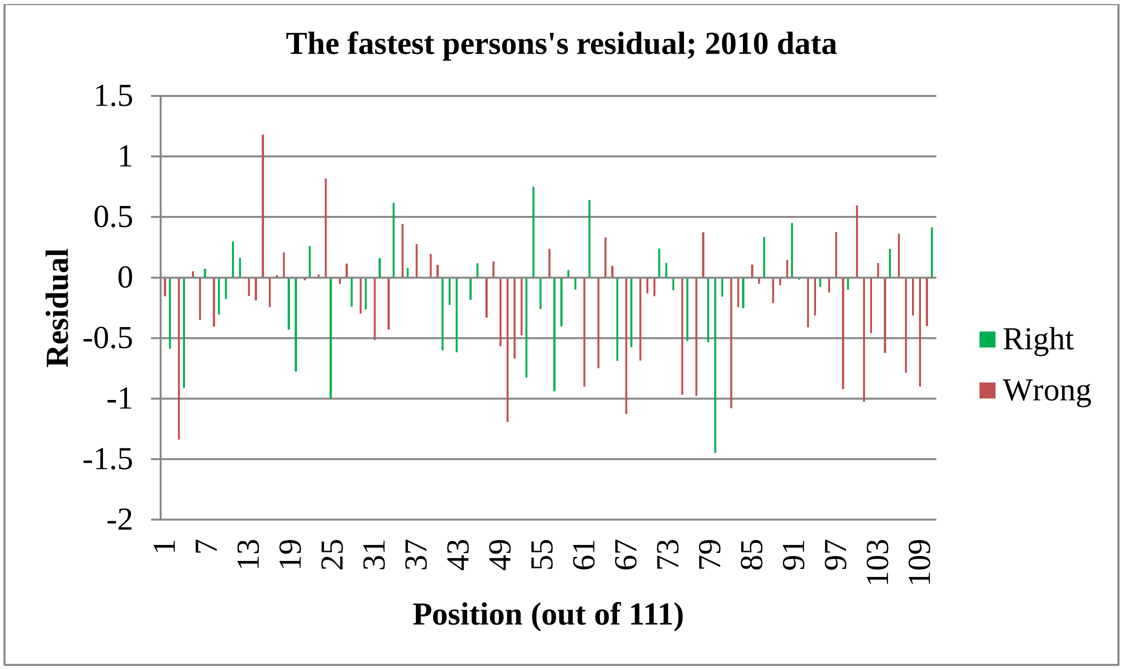


Figure 2.12

The fastest examinee's residuals



.

Figure 2.13

The slowest examinee's residuals

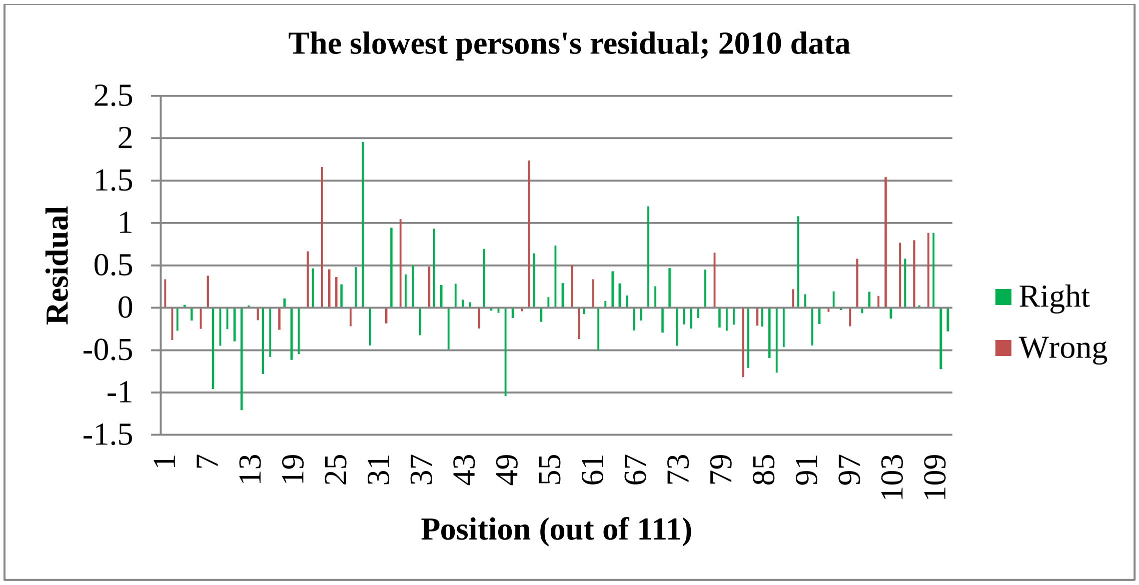


Figure 2.14

Residuals of an examinee with possible rapid guessing

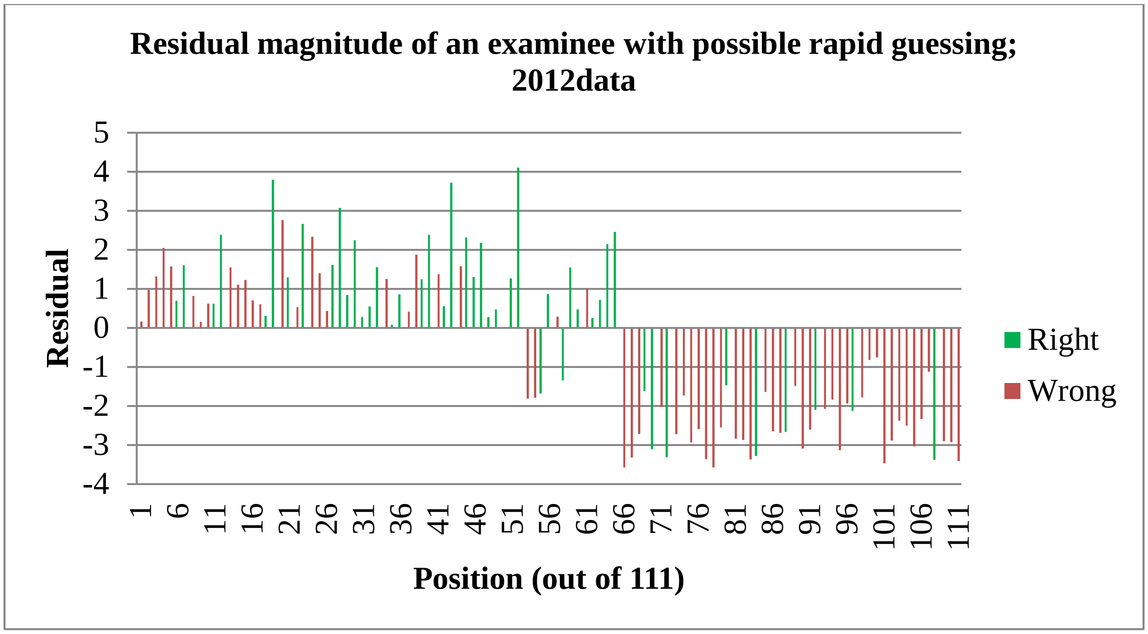


Figure 2.15

Residuals of an examinee with possible item pre-knowledge

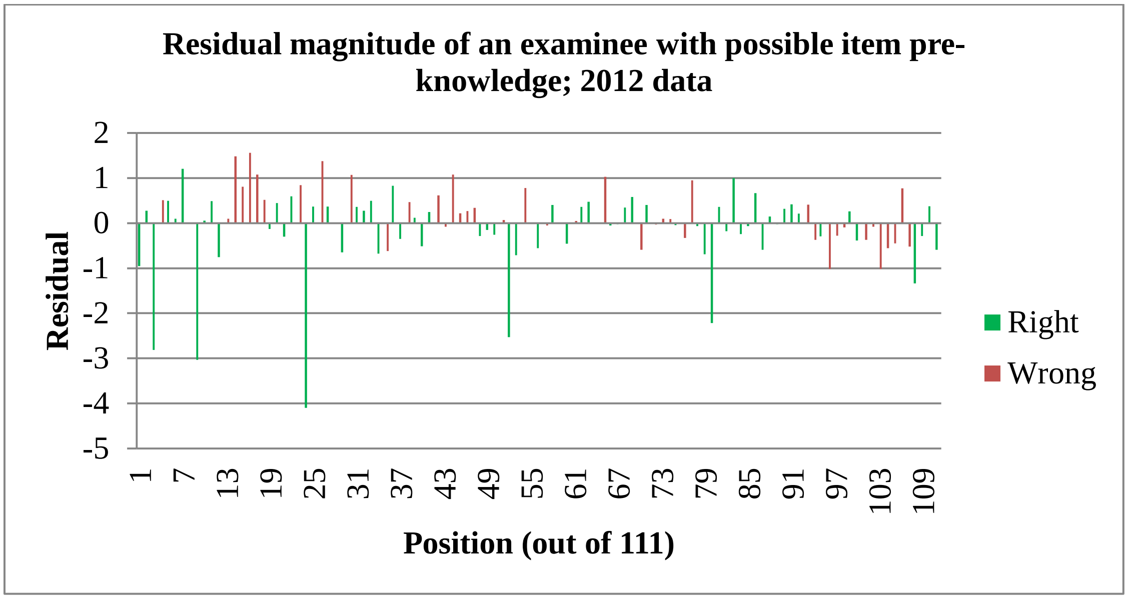


Figure 2.16

Residuals of a possibly compromised item

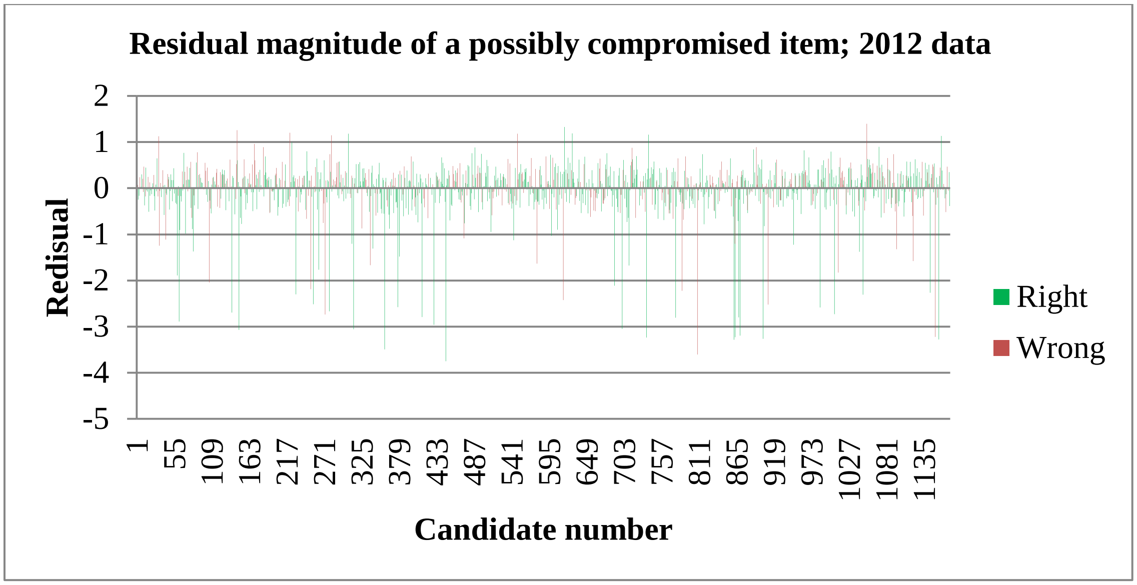
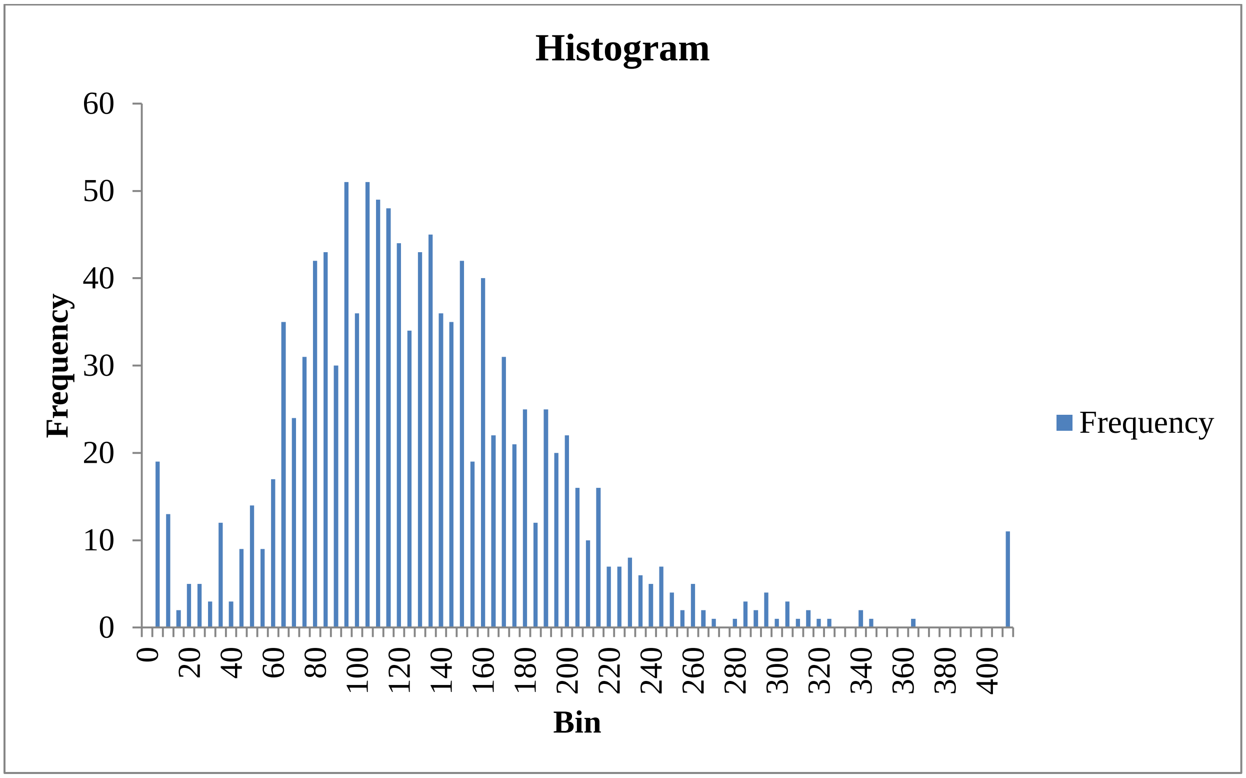


Figure 2.17

A histogram of response times for a compromised item



Appendix 2.2

Table 2.1

RMSE for lognormal distribution for each item

Item	RMSE	Item	RMSE	Item	RMSE	Item	RMSE	Item	RMSE	Item	RMSE
1	0.061456	21	0.018071	41	0.028718	61	0.036371	81	0.045725	101	0.032388
2	0.058088	22	0.012559	42	0.035687	62	0.016852	82	0.016147	102	0.014514
3	0.024868	23	0.020381	43	0.007752	63	0.026771	83	0.010686	103	0.05304
4	0.070035	24	0.066025	44	0.00647	64	0.032169	84	0.011647	104	0.013815
5	0.039921	25	0.032247	45	0.023881	65	0.045329	85	0.025101	105	0.011957
6	0.029387	26	0.022081	46	0.017511	66	0.018106	86	0.029131	106	0.032573
7	0.007846	27	0.041071	47	0.035795	67	0.050078	87	0.02573	107	0.015469
8	0.014082	28	0.061736	48	0.036802	68	0.080492	88	0.02417	108	0.022955
9	0.032429	29	0.021027	49	0.023582	69	0.01668	89	0.020015	109	0.019811
10	0.057345	30	0.01282	50	0.073618	70	0.041132	90	0.022743	110	0.051438
11	0.012221	31	0.046482	51	0.012463	71	0.026905	91	0.018491	111	0.10739
12	0.018225	32	0.023364	52	0.031049	72	0.019918	92	0.067601		
13	0.055596	33	0.016521	53	0.024465	73	0.009533	93	0.054719		
14	0.029585	34	0.038178	54	0.026229	74	0.082316	94	0.061261		
15	0.013739	35	0.022779	55	0.083564	75	0.040314	95	0.032404		
16	0.0291	36	0.024982	56	0.015755	76	0.028896	96	0.014471		
17	0.018592	37	0.03181	57	0.02386	77	0.015409	97	0.02511		
18	0.030646	38	0.044263	58	0.009845	78	0.037106	98	0.017929		
19	0.02108	39	0.013589	59	0.010936	79	0.074831	99	0.029575		
20	0.022965	40	0.013401	60	0.035583	80	0.016792	100	0.076965		

Table 2.2
RMSE for three distributions for 2010 sample

	RMSE		
Distribution	Mean	Min	Max
Lognormal	0.0317	0.0065	0.1074
Gamma	0.0482	0.0085	0.1258
Normal	0.0854	0.0616	0.1307

Table 2.3
RMSE for three distributions for 2012 sample

	RMSE		
Distribution	Mean	Min	Max
Lognormal	0.0607	0.0094	0.2074
Gamma	0.0719	0.0139	0.2149
Normal	0.1093	0.0307	0.2441

Table 2.4
Paired t-test results

Paired Samples Statistics

		Mean	**N**	**Std. Deviation**	**Std. Error Mean**
Pair1	Year_2010	1.47	111	1.731	.164
	Year_2012	1.93	111	1.939	.184

Paired Samples Test

	Paired Differences							
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig.(2-tailed)
				Lower	Upper			
Pair1 Year_2010 – Year_2012	-.459	1.645	.156	-.769	-.150	-2.943	110	.004

Table 2.5

Detection rate for simulated studies

# of items #of examinees	**1**	**5**	**10**	**30**	**50**	**100**
10	.7400	.7000	.6460	N/A	N/A	N/A
50	.7360	.6608	.6244	N/A	N/A	N/A
100	.6940	.6638	.6372	.4567	.2682	.0192

Chapter 3. Using response time to increase measurement efficiency in computerized adaptive testing

The shift in test delivery from paper-and-pencil testing to computerized adaptive testing was motivated by an important goal: increasing measurement efficiency, which means improving ability estimation accuracy with the same length of test or shortening the test while maintaining the same level of accuracy (Lord, 1980; Meijer & Sotaridona, 2006). Computerized adaptive tests (CATs) have long been proved to give good estimates of examinee proficiency with fewer numbers of items than fixed length tests that cover the same content (Lord, 1980). For example, a current large-scale licensure test using CAT reports good ability estimation accuracy. According to a simulation study, the correlation between true ability and estimated ability is 0.924, and the classification accuracy is 93.7%[7]. But it would be better if we could improve classification accuracy even by 1% given the large volume who take this test (there are more than 50,000 examinees in three months)[8].

At the same time, the test length for the examinees whose ability is near the cut score is usually very long in variable length adaptive tests that have stopping rules related to decision making. That is the CAT algorithm needs more items to determine whether the examinee's ability is above the cut score or below the cut score with the same confidence than it would need for other examinees whose ability is far above or below the cut score. For example, in the current large-scale licensure test, about 10% of examinees finish the test with the maximum number of

[7] It means that 93.7% of examinees will be classified as pass when they should pass or classified as fail when they should fail. About 6.3% of examinees will be classified as pass when they do not deserve pass or as fail when they should pass.

[8] If I can improve 1% of classification accuracy, it will make more than 500 examinees getting correct classification in three months.

items, which is 250 items, while other examinees can finish the test with as few as 60 items. If we can reduce the test length for these examinees while keeping the same level of accuracy, it will benefit the testing program as well as the examinees[9].

Most current CAT algorithms estimate examinee proficiency based on the item score vector[10]. In this proposed study, I want to explore how to improve ability estimation and shorten the test with the response time for each item as collateral information in CAT procedures. The reason why using response time data can increase measurement efficiency is that information can be borrowed from response times to estimate ability based on the correlation between persons' speed and ability. The larger the correlation is, the more benefit there would be.

The research questions for this essay include:

- To what extent is ability estimation improved with response times as collateral information in a CAT procedure?
- To what extent is test length shortened for examinees whose abilities are near the cut score with response times as collateral information in a CAT procedure?

In the first section of this essay, I review the research literature on using response times to select items and improve ability estimation. The second section describes the models I used in the study. The third section introduces my research design. In the fourth section, I present the results from simulation studies.

[9] For the examinees, they can save time, money (test fee is partially based on the test length) and energy. For the testing program, they do not need to develop a lot of items if the ability can be estimated with shortened length of test with the same level of accuracy. Also more examinees can take the test within the certain account of time since some of them will not occupy the computer for a long time.

[10] If an examinee got an item correct, he/she will get a higher ability estimate no matter how much time he/she took.

Literature review

In this section, I review research on using response times to select items and improve ability estimation. Several researchers have tried to use response times to improve item selection and ability estimation. van der Linden (2008) used response times for item selection with the correlation between ability and speed at .2, .4, .6, and .8. The simulation studies showed that the use of response times was profitable, especially for shorter tests. For example, the mean squared error for a test of length n=10, with the use of response times for correlation .6 was virtually identical to the mean squared error for a test of double length without the use of response times. Also, the gains were larger for the extreme θs since the item pool did not have a lot of very easy or hard items to measure the ability of very low- or high-ability examinees. The results also showed that a small correlation of .2 was already enough to have a large impact.

van der Linden, Klein, Entink, and Fox (2010) used response times as collateral information to improve the estimation of the person and item parameters. In addition to the regular information in the response data, two sources of collateral information are identified: (a) the joint information in the responses and the response times summarized in the estimates of the second-level parameters[11] and (b) the information in the posterior distribution of the response parameters given the response times. The study showed that the use of this posterior predictive distribution both improved the accuracy and reduced the bias of parameter estimates.

Wang and Jiao (2011) used response times, student background information (such as gender and ethnicity), and academic information (test scores from other examinations) as collateral information to improve the precision of parameter estimates. The results showed that in general, models with collateral information fit better than models without any collateral information.

[11] Please refer to the models presented in Chapter 2 on pages 54-55.

Therefore, the contribution of collateral information in improving parameter estimation should not be ignored.

Since among testing agencies there exists a general hesitation to report scores that are calculated from any information other than the performances of the test takers on the test (van der Linden, 2008), some researchers only use the response times for item selection. Fan, Wang, Chang, and Douglas (2012) utilized response time distributions for item selection in CAT. Because traditional methods for item selection in CAT only focus on item information without taking into account the time required to answer an item, some examinees may receive a set of items that take a very long time to finish, and information is not accrued as efficiently as possible. Fan et al. (2012) proposed two item-selection criteria that utilize information from a lognormal model for response times. The first modifies the maximum information criterion to maximize information per time unit. The second is an inverse time-weighted version of *a*-stratification[12] that takes advantage of the response time model, but achieves more balanced item exposure than the information-based techniques. The results showed that when response time information was taken into account, the average time for examinees to complete a test was reduced.

[12] Some items are good at differentiating examinees' abilities, which means high-ability examinees will answer these items correctly while low-performance examinee will answer these items incorrectly. The term used to quantify such item characteristic is discrimination parameter or α. An item with high α means high discrimination power. At the beginning of CAT, the computer does not know a lot of an examinee's true ability, so we do not want the computer to select high-α item first. Usually we will order the items according to their α into high- α, middle-α, and low- α strata and let the computer selects items from middle- α items first. Later when the computer knows more about an examinee's ability, it can choose items from high- α stratum and uses these items more efficiently. We may not want the computer selects items from low- α stratum depending on whether we have enough items in other two strata. This is call α-stratification method.

The previous research showed that response times can potentially provide a wealth of information when exams are administered by computer. Therefore, I used response times to attempt to improve ability estimation for a large-scale licensure test.

Models used

The models used in this essay are the same as the models described in the previous essay.

Research design

To explore whether using response times as collateral information can increase measurement efficiency, I conducted two simulation studies. The item pool for the simulation study was an actual item pool from the large-scale licensure test. Because this was a simulation study, no actual items were needed. Only a file with the difficulty parameters, time intensity parameters and time discrimination parameters for the items from the item pool were required so there is no concern about the security of the item pool.

The first simulation program that I developed implements the process used by the large-scale licensure test including the item selection rule, the exposure control procedure, and the stopping rule relative to the current cut-score for passing the test. This program simulated 51,480 examinees using mean and standard deviation from the empirical data to set a base rate for the correlation between true and estimated proficiency, the proportion of examinees that are correctly classified and mean test length for these examinees.

Then a second simulation study with a new CAT procedure incorporating response time model was developed while all other processes were the same as the original test. This was done with a random sample of examinees from the examinee population using the same mean and standard deviation as in the first simulation program. The results from this analysis were then compared to the base rate analysis in terms of the correlation between true and estimated proficiency, the proportion of examinees that are correctly classified and mean test length for these examinees.

Results

Results for simulation 1

The first simulation checked the function of the CAT for a random sample of 51,480 examines sampled from a normal distribution with a mean of .15 and a standard deviation of .59. These data are the empirical statistics from the test administration for the large-scale licensure examination. The distribution of sampled true θs is in Figure 3.1:

For simulation study, these 51,480 true θs are input for CAT program. The output is 51,480 estimated θs. The distribution of estimated θs is in Figure 3.2:

From Figure 3.1 and Figure 3.2, we can see that there are some differences between the true θs and estimated θs. In order to compare the true θs and estimated θs, I plotted the true θs on x axis and the estimated θ on y axis. The scatter plot of the true θs and estimated θs is in Figure 3.3.

In Figure 3.3, the points on the identity line mean the true θs and the estimated θs are exactly the same. Other points are close to the identity line, which means the true θs and estimated θs are similar. Actually, the correlation between the true θs and estimated θs is 0.9308, which means they are highly correlated with each other, indicating that the estimation is working appropriately. Please note that the shape around (-0.16, -0.16) is unusual, which has much narrower dispersion than other places. The reason is that for examinees whose ability is near the cut score (-0.16), the CAT algorithm cannot determine whether their ability is above or below the cut score with 95% confidence with a short test, so the algorithm continues giving items to these examinees near the cut score, which results in a long test for them. With a longer test, the amount of error of measurement is small, so there is less spread for the scatter plot around the cut score. This can also explain the unusual distribution near -0.16 in Figure 3.2.

Since we know examinees' true abilities in the simulation study, and we have a specific cut score, we know of 51,480, how many examinees should pass and how many should fail. After obtaining the estimated abilities, we can also classify them as pass or fail according to the same cut score. Table 3.1 shows the classification accuracy for the simulation using the regular CAT program.

In Table 3.1, 34,397 examinees passed the exam when they should have passed according to their true abilities and 13,849 failed the exam when they should have failed because their true abilities were below the cut score. At the same time, 1,754 examinees failed the exam when they should have passed and 1,480 passed the exam when they should have failed according to their true abilities. These are misclassifications. So the classification accuracy is (34,397+13,849)/51,480 = 93.72%.

From Table 3.1, we can see that this large-scale licensure exam has good classification accuracy. However, a related issue is test length. In general, longer tests tend to have better classification accuracy. The number of items administered to the 51,480 examinees is shown in Figure 3.4.

This large-scale licensure exam is a variable-length CAT test. Therefore, the test length for each candidate can be different. From Figure 3.4, we can see the minimum test length is 60 items while the maximum length is 250 items. This is decided by the CAT stopping rule. For such a large-scale test, there are two stopping rules related to test length for a simulation study. (a) 95% Confidence Interval Rule. After 60 items, the computer sill stop administering items when the 95% confidence interval around the current ability estimate does not include the cut score, (b) Maximum-Length Exam Rule. When your ability is very close to the passing standard, the computer continues to give you items unstill the maximum number of items is reached, which is 250 items. The average number of items taken by 51,480 examinees is 104.0341.

In order to examine variations in terms of (a) the correlation between true ability and estimated ability, (b) classification accuracy, and (c) average number of items for different random samples (with same mean, standard deviation, and sample size), I generated four random samples. The results are showed in Table 3.2.

From Table 3.2, there is not a lot of variation across the samples. The average correlation between true ability and estimated ability is 0.93. The classification accuracy is 0.94. The average number of items taken by 51,480 examinees is 104 items.

Results for simulation 2

Before showing the results for simulation 2, I need to describe how we can use response times as collateral information to improve the estimation of ability in an IRT model. van der Linden et al. (2010) made a good analogy of using a Bayesian model prior to improve the ability estimation. In their article (p.329), they wrote:

> Suppose the test takers are from a population with a normal distribution of ability $N(\mu_\theta, \sigma_\theta^2)$, of which the mean μ_θ and variance σ_θ^2 have already been estimated with enough precision to treat them as known. Estimates of θ_j that capitalize on this information should be based on the posterior distribution
>
> $$f(\theta_j \mid u_j, \mu_\theta, \sigma_\theta^2) \propto f(u_j \mid \theta_j) f(\theta_j \mid \mu_\theta, \sigma_\theta^2), \qquad (1)$$
>
> Where $u_j = (u_{1j}, \ldots, u_{nj})$ are the responses by test taker *j* on the *n* items in the test and $f(u_j \mid \theta_j)$ is the probability of the observed response vector by the test taker under the response model.
>
> The mean of this posterior distribution, which is often used as a point estimate of θ_j, is generally known to have a smaller mean square error than a classical estimate based on the probability of the observed data, $f(u_j \mid \theta_j)$ only. The decrease is due to the information in the population density $f(\theta_j \mid \mu_\theta, \sigma_\theta^2)$ in the right-hand side of

Equation 1, which shows, for instance, where the ability parameters in the population are concentrated and how much they are dispersed. The decrease is compensated for by an increase in the bias of the ability estimate toward the mean of the population of test takers or the domain of items.

The same principle as in Equation 1 is demonstrated on page 333 in their article but this time for a test taker *j* with response vector $\mathrm{u}_{\mathrm{j}} = (\mathrm{u}_{1\mathrm{j}}, \ldots, \mathrm{u}_{\mathrm{nj}})$ and response time vector $t_{\mathrm{j}} = (t_{1\mathrm{j}}, \ldots, t_{\mathrm{nj}})$ assuming that mean $\mu_{\theta\tau}$ and covariance matrix $\Sigma_{\theta\tau}$ have already been estimated during item calibration.

$$f(\theta_{\mathrm{j}} \mid \mathrm{u}_{\mathrm{j}}, \mathrm{t}_{\mathrm{j}}, \mu_{\theta\tau}, \Sigma_{\theta\tau}) \propto f(\mathrm{u}_{\mathrm{j}} \mid \theta_{\mathrm{j}}) f(\theta_{\mathrm{j}} \mid \mathrm{t}_{\mathrm{j}}, \mu_{\theta\tau}, \Sigma_{\theta\tau}), \qquad (2)$$

The result has a simple form that is entirely analogous to Equation 1. It shows that when the response times are used as collateral information, θ_{j} is estimated from the probability $f(\mathrm{u}_{\mathrm{j}} \mid \theta_{\mathrm{j}})$ of the observed response vector u_{j} as when the response times are ignored but with the original prior distribution of θ in Equation 1 replaced by the conditional posterior distribution of θ_{j} given the response times t_{j} for the test taker.

The result in Equation 2 helps identify three different sources of information about θ_{j}:

1. The information directly available in u_{j} in the first factor of Equation 2, that is, the regular model probability $f(\mathrm{u}_{\mathrm{j}} \mid \theta_{\mathrm{j}})$ associated with the observed response vector.
2. The information summarized in the estimates $\mu_{\theta\tau}$ of and $\Sigma_{\theta\tau}$ in the second factor. This information is derived from the vectors of responses and RTs in the entire sample of test takers. These estimates generalize the role of those of μ_{θ} and σ_{θ}^{2} in Equation 1.
3. The information in the shape of the conditional posterior distribution of the response parameters given the response times. Unlike the estimates of the population parameters in the preceding source of information, the information in this vector is unique for each individual test taker.

Procedures for using response times as collateral information for ability estimation in this study:

(1) Generate true ability and true speed according to empirical statistics. For this study, I generate ability and speed from multivariate random normal distribution with mean= [0 0] and covariance matrix [1.00 0.1; 0.1 0.1]. According to this covariance matrix, the empirical correlation between ability and speed is 0.3162.

(2) Create response time for each person on each item according to person speed, time intensity parameter, and time discrimination parameter and response time model.

(3) Estimate person speed after each item according to response time string and item parameters using maximum likelihood estimation.

(4) According to the covariance matrix, work out the regression model with ability as dependent variable and speed the independent variable. For this empirical covariance matrix, the regression model happens to be ability=speed with slope of 1 and intercept of 0. The standard error of estimation for the regression prediction is the standard deviation of (speed) * $\sqrt{1-r^2}$, where *r* is the correlation between ability and speed. For this study, the standard error of estimation for regression is 0.9499.

(5) Calculate the likelihood of the observed response vector multiply the normal distribution with mean of estimated speed and standard deviation of 0.9499. Then find θ that maximizes the product. This is the estimated θ using response times as collateral information. Procedures (2), (3), (4), and (5) are repeated after administrating each item.

The results for simulation 2 are showed in Table 3.3.

Comparing Table 3.3 to Table 3.2, the correlations between true ability and estimated ability decreased to a large degree when using response times as collateral information to estimate ability. Classification accuracy also dropped a lot. The only promising result is that the average number of items taken by examinees was reduced to a large degree, but at the cost of less accuracy of ability estimation. The usefulness of response times for ability estimation depends on the correlation between person ability and person speed. In the next section, I present the results for the correlation between person ability and speed of 0.8 and 0.95 respectively in Table 3.4 and Table 3.5.

From Table 3.3 to Table 3.4 and from Table 3.4 to Table 3.5, with the increase of the correlations between person ability and person speed, the correlations between true ability and estimated ability, classification accuracy, and average number of items taken by examinees also increased. However, even with the correlation of 0.95, which is not realistic for actual tests, the correlations between true ability and estimated ability and classification accuracy cannot match the base rate. The results from simulation 2 showed that using response times as collateral information does not improve ability estimation for this large-scale licensure examination using the specified estimation procedure.

Discussions and conclusions

The purpose of this study is to explore whether ability estimation is improved with response times as collateral information in a CAT procedure for a large-scale licensure examination. The results from this study showed that for this particular test, using response times does not improve ability estimation in terms of the correlations between true ability and estimated ability and classification accuracy. There are several possible reasons for these results. First, the original test is high quality as shown by the 0.93 correlations between true ability and estimated ability and the 0.94 classification accuracy with average test length of 104 items. van der Linden’s (2008) research showed that the mean squared error for a test of length n=10, with the use of response times for correlation .6 were virtually identical to the mean squared error for a test of double length without the use of response times. But the gains are reduced with the test length. For a test length of 104 items, it is very difficult to improve ability estimation with other information. Second, the variance of speed for the population who took this large-scale licensure test is very small (0.1), which led to a lot of shrinkage on the θ estimation by regression effect. The shrinkage reduces the variations of estimated ability and thus reduces the correlation between the true ability and estimated ability because of restricted range. Therefore, if we use the same procedure with a different population with large variations of speed, the results may be better.

The third possibility is related to the estimation procedure used in this study. For this research, speed is estimated first, and then used to estimated ability, which may increase the ability estimation bias and lead to more misclassification. A better way is to estimate speed and ability at the same time, which can be done by Markov chain Monte Carlo (MCMC) estimation. Actually, when I applied MCMC techniques to estimate the final θ using the response string and response time string for the whole test, the final estimation is highly correlated to the estimation using response time only. Figure 3.5 showed the scatter plot for the ability estimation using response only and using response and response times.

From Figure 3.5, we can see that two estimations are very similar. Actually, the correlation between these two sets of estimation is 0.9551. Therefore, if we can use MCMC estimation in a CAT procedure, we should get results that are similar to (if not better than) the results from the original CAT only using item responses. But one of the primary drawbacks of MCMC is its heavy computational demand (Kim & Bolt, 2007). The sampling procedures that underlie the MCMC methodology generally require a very large number of iterations before model parameters can be reliably estimated. Therefore, it is not uncommon for a single estimation run to take several hours, or even a day or more. Therefore, it is very difficult to use this for a CAT procedure. For CAT, estimation needs to be done after each item is administered, and then the next item is chosen according to the newly estimated ability with difficulty parameter matching the ability estimation. But it is not possible for examine to wait for a long time to see the next item. That’s why I chose to use the approximation procedure, but this procedure cannot improve the estimation even though it is fast enough to be used in CAT.

There are several implications from this research. First, it does not always work to apply theoretically promising methodology to a practical examination. For this study, the usefulness of response times is limited by the population (small variance of speed) and the nature of CAT (which requires fast estimation). Actually, there is not enough research connecting the theoretical research to the actual world of testing. In order to increase the application of theoretically sound methodology to practice, more research like this should be done. Second, in order to use response time as collateral information to improve ability estimation in CAT, a new estimation method is needed which has less estimation bias and is fast enough to be able to choose the next item without causing unreasonable delays to test takers.

APPENDICES

Appendix 3.1

Figure 3.1

Distribution of sampled true θs

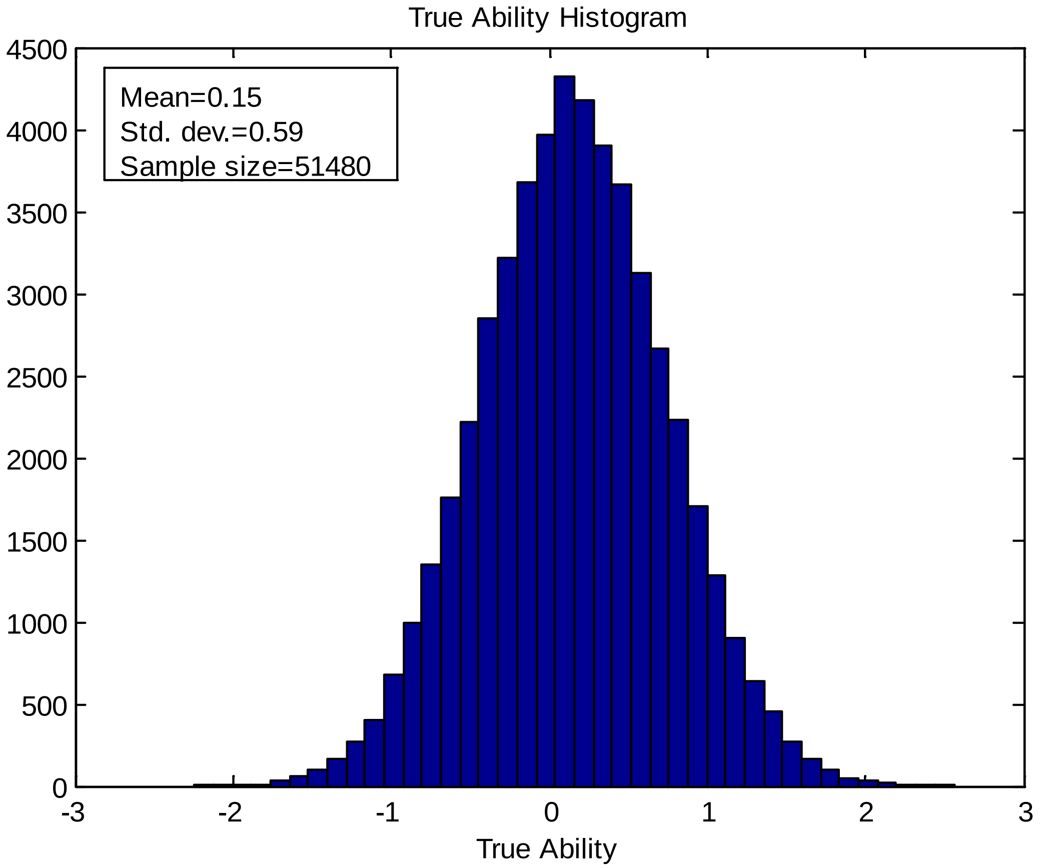


Figure 3.2

Distribution of estimated θs

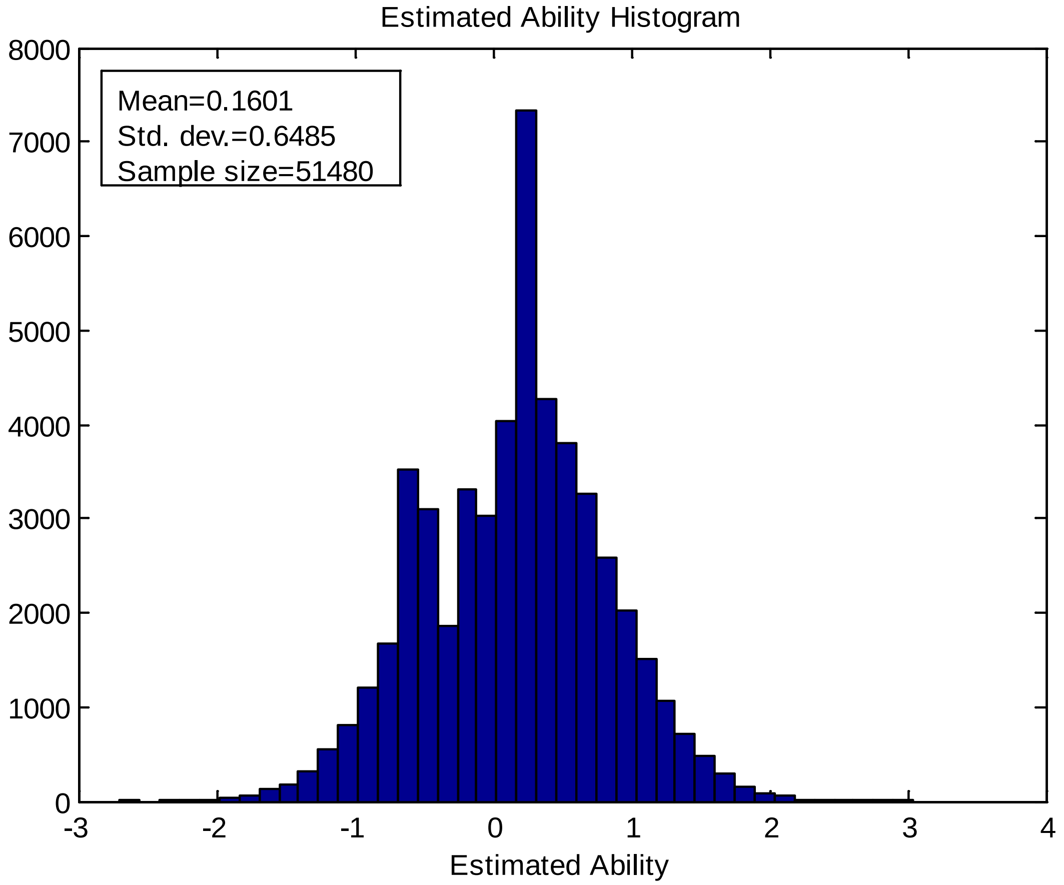


Figure 3.3

Scatterplot of true θs and estimated θs

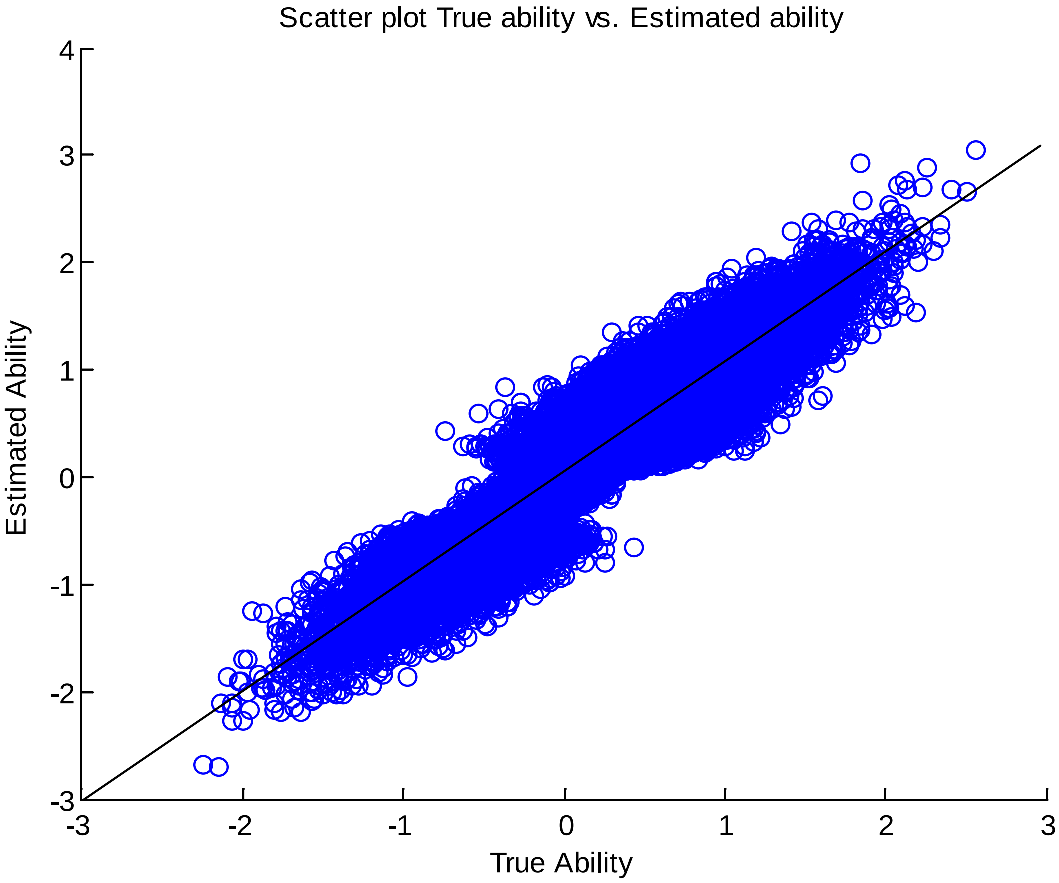


Figure 3.4

Distribution of test lengths for 51,480 examinees

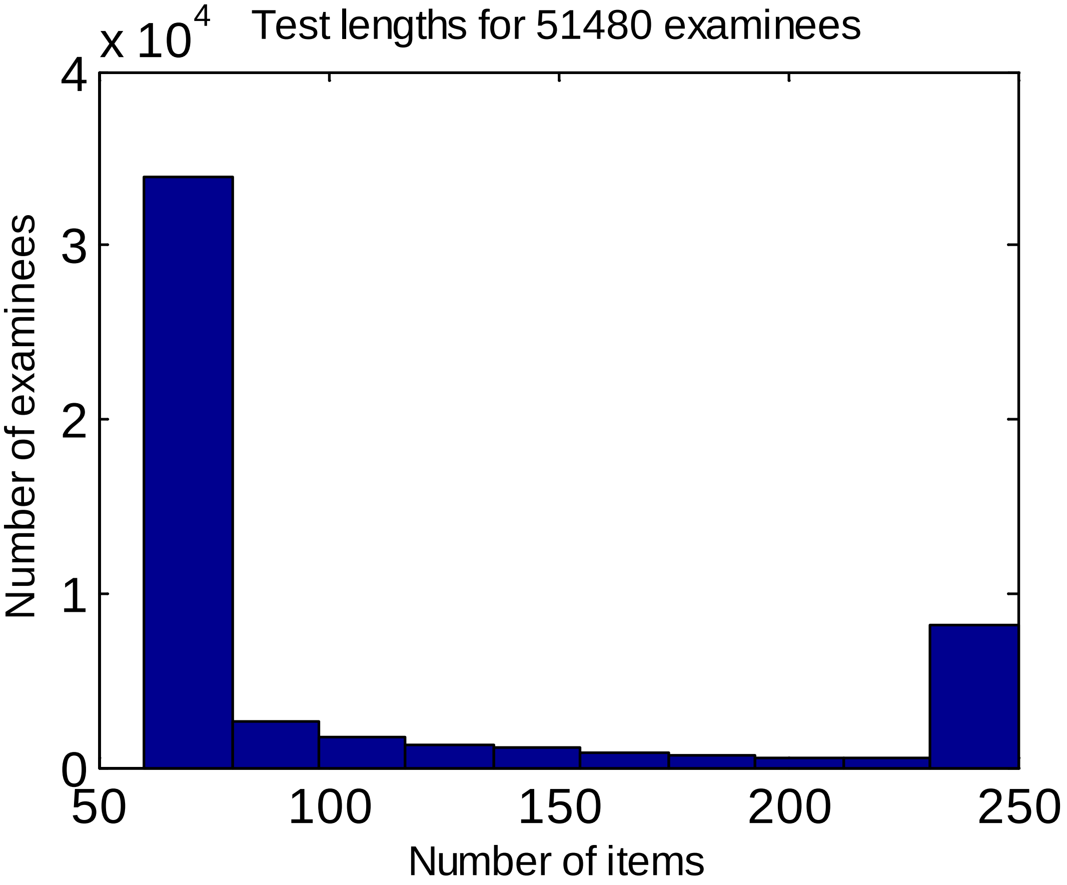
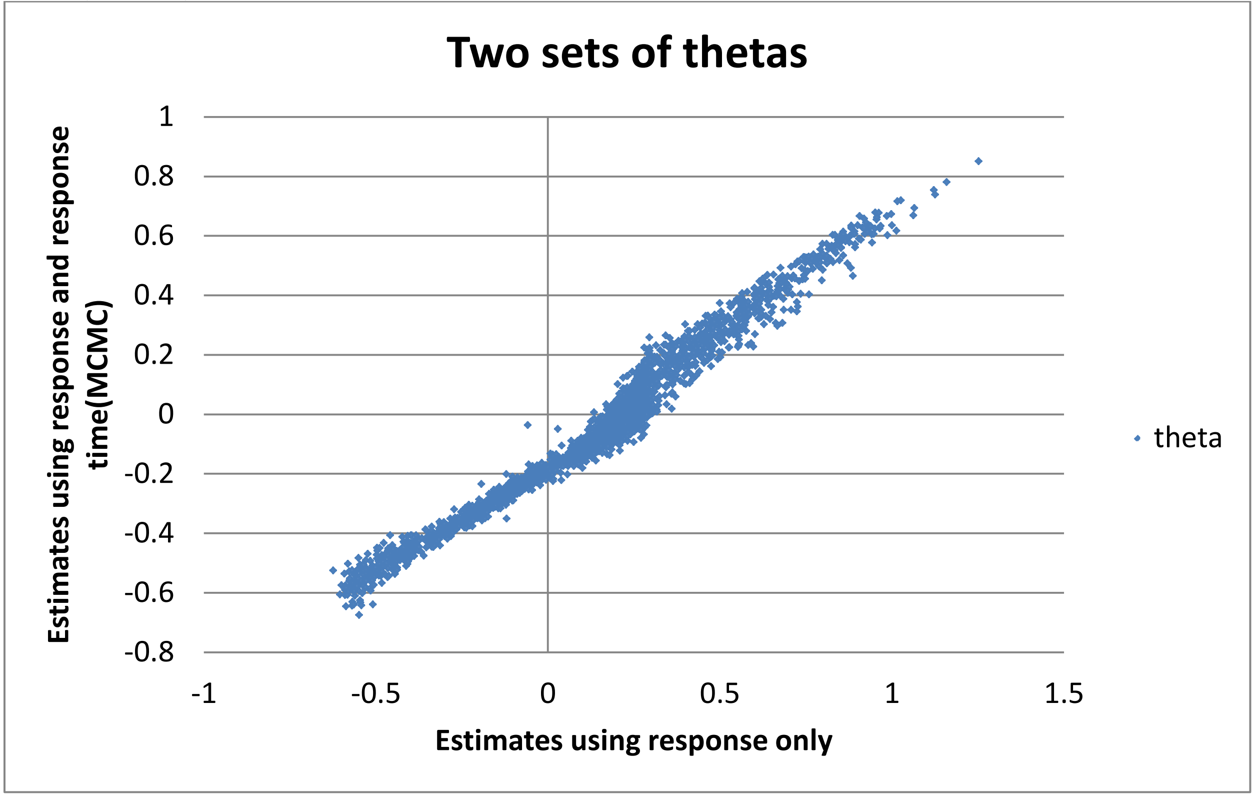


Figure 3.5

Scatter plot of final ability estimation using response only and using response and response times (MCMC)



Appendix 3.2

Table 3.1

Classification accuracy for 51480 examinees using response data only

		Estimated θ		
		Pass	Fail	Total
True θ	Pass	34397	**1754**	36151
	Fail	**1480**	13849	15329
	Total	35877	15603	51480

Table 3.2

Base rate results for 4 random samples

	Correlations between true ability and estimated ability	Classification accuracy	Average number of items taken by 51480 examinees
Random sample 1	0.9308	0.9372	104.0341
Random sample 2	0.9299	0.9389	103.8966
Random sample 3	0.9307	0.9379	103.5735
Random sample 4	0.9312	0.9406	103.1302
Average	0.93065	0.93865	103.6586

Table 3.3

Results from simulation 2 for 5 random samples using response and response time data (correlation between ability and speed= 0.3)

	Correlations between true ability and estimated ability	Classification accuracy	Average number of items taken by 51480 examinees
Random sample 1	0.3207	0.629	61.979
Random sample 2	0.2527	0.67	62.12
Random sample 3	0.4374	0.69	61.88
Random sample 4	0.3138	0.62	62.02
Random sample 5	0.3436	0.63	62.19
Average	0.3336	0.6478	62.0738

Table 3.4

Results from simulation 2 for 4 random samples using response and response time data (correlation between ability and speed = 0.8)

	Correlations between true ability and estimated ability	Classification accuracy	Average number of items taken by examinees
Random sample 1	0.7299	0.78	68.3
Random sample 2	0.6745	0.69	65.96
Random sample 3	0.6471	0.71	68.29
Random sample 4	0.6638	0.74	65.68
Average	0.6788	0.73	67.0575

Table 3.5

Results from simulation 2 for 3 random samples using response and response time data (correlation between ability and speed = 0.95)

	Correlations between true ability and estimated ability	Classification accuracy	Average number of items taken by examinees
Random sample 1	0.7794	0.833	68.676
Random sample 2	0.8	0.843	69.541
Random sample 3	0.7825	0.82	69.565
Average	0.7873	0.832	69.261

REFERENCES

REFERENCES

Akiba, M., LeTendre, G., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries around the world. *Educational Researcher*, 36(7), 369-387.

Allen, M. (2003). *Eight questions on teacher preparation: What does the research say?* Denver, CO: Education Commission of the States. Retrieved from http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED479051

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Tsai, Y-M. (2010). Teachers' mathematicsematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133-180.

Begle, E.G. (1979). *Critical variables in mathematics education: Findings from a survey of the empirical literature.* National Council of Teachers of Mathematics, Reston, Virginia.

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit der Lehrerbildung [Empirical findings for the effectiveness of the teacher formation]. In S. Blömeke, P. Reinhold, G. Tulodziecki, & J. Wildt (Eds.), *Handbuch lehrerbildung* [Manual teacher formation] (pp. 59-91). Bad Heilbrunn/Braunschweig, Germany: Klinkhardt/Westermann.

Bloemeke, S., Paine, L., Houang, R.T., Hsieh, F.-J., Schmidt, W.H., Tatto, M.T., Bankov, K., Cedillo, T., Cogan, L., Han, S.I., Santillan, M., & Schwille, J. (2008). Future teachers' competence to plan a lesson: First results of a six-country study on the efficiency of teacher education. *ZDM–The International Journal on Mathematics Education, 40*(5), 749-762.

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education*, 62(2), 154-171.

Boyd, D., Grossman, P. Lankford, H. Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*(4), 416-440.

Briscoe, C., & Stout, D. (1996). Integrating mathematics and science through problem centered learning in methods courses: Effects on prospective teachers' understanding of problem solving. *Journal of Elementary Science Education*, 8(2), 66-87.

Clift, R.T., & Brady, P. (2005). Research on methods courses and field experiences. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying teacher education*: *The report of the AERA Panel on Research and Teacher Education* (pp.309-424).Washington, DC: American Educational Research Association.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6), 673-682.

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education.* Mahwah, NJ: Lawrence Erlbaum.

Cohen, A.S. & Wollack, J.A. (2006). Test Administration, Security, Scoring and Reporting. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355-386). New York: Macmillan.

Croninger, R.G., Rice, J.K., Rathbun, A., & Nishio. M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review*, 26(3), 312-324.

Fan, Z., Wang, C., Chang, H., & Douglas, J. (2012) Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, 37 (5), 655-670.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, 20(7), 1-14.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://ncctq.learningpt.org/publications/LinkBetweenTQ andStudentOutcomes.pdf

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

Grossman, P., Compton, C., Igra, D., Ronfeldt, M.,. Shahan, E., & Williamson, P.W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055-2100.

Hill, H.C. (2010). The nature and predictors of elementary teachers' mathematicsematical knowledge for teaching. *Journal for Research in Mathematics Education*, *41*(5), 513-545.

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

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers’ mathematicsematical knowledge: What knowledge matters and what evidence counts? In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp.111-155). Charlotte, NC: Information Age.

Ingvarson, L., Beavis, A., & Kleinhenz, E. (2007). Factors affecting the impact of teacher education courses on teacher preparedness: Implications for accreditation policy. *European Journal of Teacher Education*, 30(4), 351-381.

Kaplan, L. S., & Owings, W. A. (2001). Teacher quality and student achievement: Recommendations for principals. *NASSP Bulletin*, 85(628), 64-73.

Kim, J., & Bolt, D.M. (2007). Estimating item response theory models using Markov Chain Monte Carlo Methods. *Educational Measurement: Issues and Practice*, 38-51.

Kim, M.K., & Sharp, J. (2000). Investigating and measuring preservice elementary mathematics teachers' decision about lesson planning after experiencing technology-enhanced methods instruction. *Journal of Computers in Mathematics and Science Teaching, 19*(4), 317-338.

Konig, J., Bloemeke, S., Paine, L., Schmidt, W.H., & Hsieh, F.-J. (2011). General pedagogical knowledge of future middle school teachers: On the complex ecology of teacher education in the United States, Germany, and Taiwan. *Journal of Teacher Education, 62*(2), 188-201.

Langrall, C.W., Thornton, C.A., Jones, G.A., & Malone, J.A. (1996). Enhanced pedagogical knowledge and reflective analysis in elementary mathematics teacher education. *Journal of Teacher Education, 47*(4), 271-282.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mcleod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, 23, 147-160.

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8(3), 261-272.

Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing*. Law School Admission Council Computerized Testing Report, 03-03.

Mewborn, D.S. (1999). Reflective thinking among preservice elementary mathematics teachers. *Journal for Research in Mathematics Education, 30*(3), 316-341.

Monk, D. H., & King, J. A. (1994). Multilevel teacher resource effects on pupil performance in secondary mathematics and science: The case of teacher subject-matter preparation. In R. Ehrenberg (Ed.), *Choices and consequences: Contemporary policy issues in education* (pp. 29-58). Ithaca, NY: ILR Press.

Mullens, J.E., Murnane, R.J., & Willett, J.B. (1996). The contribution of training and subject matter knowledge to teaching effectiveness: A multilevel analysis of longitudinal evidence from Belize. *Comparative Education Review*, 40(2), 139-157.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future. Report of the National Commission on Teaching & America's Future.*Woodbridge, VA: Author.

Rice, J. K. (2003). *Teacher quality Understanding the effectiveness of teacher attributes.* Washington, DC: Economic Policy Institute.

Rowan, Chiang, B. F., & Miller, R.J. (1997).Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, 70, 256-284.

Schmidt, W.H., Bloemeke, S., & Tatto, M.T. (2011). *Teacher education matters: A study of middle school mathematics teacher preparation in six countries.* New York: Teachers College Press.

Schmidt, W.H., Cogan, L., & Houang, R. (2011). The role of opportunity to learn in teacher preparation: An international context. *Journal of Teacher Education*, 62(2), 138-153.

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-22.

Smith R.W., & Davis-Becker, S.L. (2011). *Detecting suspect examinees: An application of differential person functioning analysis.* The annual conference of the National Council on Measurement in Education. New Orleans, LA.

Snijders, Tom A. B., und Bosker, Roel J., (2000). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling.* London, Thousand Oaks: Sage.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., Reckase, M. (2012). *Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries.* East Lansing: Teacher Education International Study Center, College of Education, Michigan State University.

Vacc, N.N., & Bright, G.W. (1999). Elementary preservice teachers' changing beliefs and instructional use of children's mathematicsematical thinking. *Journal for Research in Mathematics Education, 30*(1), 89-110.

Veerkamp, W. J. J. (1996). *Statistical methods for computerized adaptive testing*. Unpublished doctoral dissertation, University of Twente, The Netherlands.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics,* 31, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287-308.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73, 365-384.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34, 327-347.

van der Linden, W. J.,& van Krimpen-Stoop, E. M.L.A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, 68, 251-265.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *New developments in computerized adaptive testing: Theory and practice* (pp. 201–219). Boston: Kluwer-Nijhoff Publishing.

Wang, S., & Jiao, H. (2011). *Incorporating person covariates and response times as collateral information to improve person and item parameter estimations*. The annual meeting of the National Council on Measurement in Education. New Orleans, Louisiana.

Wayne, A.J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1), 89-122.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations.* Seattle, WA: Center for the Study of Teaching and Policy. Retrieved September 30, 2009, from http://depts.washington.edu/ctpmail/PDFs/TeacherPrep-WFFM-02-2001.pdf

Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163-183.

Youngs, P., & Qian, H. (2003). The influence of university courses and field experiences on Chinese Elementary candidates' mathematicsematical knowledge for teaching. *Journal of Teacher Education*, 64(3), 244-261.