CHEN, FANG. Ph.D. Differential Language Influence on Math Achievement. (2010). Directed by Dr. Micheline Chalhoub-Deville. 175pp.

New models are commonly designed to solve certain limitations of other ones. Quantile regression is introduced in this paper because it can provide information that a regular mean regression misses. This research aims to demonstrate its utility in the educational research and measurement field for questions that may not be detected otherwise. Quantile regression is appropriate when the assumption of a normal distribution of the error term is violated. It is most useful when the interest is at various locations along the complete distribution rather than just the central tendency.

The first part of this research used quantile regression to explore a changing relationship between language proficiency and math achievement. Results reveal that language proficiency affects math achievement differently at different math ability levels. Other commonly used covariates such as socioeconomic status and gender are also related to math achievement differently at different locations on the math score distribution. It is shown that regular mean regression analyses fail to capture this information.

The second part of the research models math growth longitudinally adjusting for language proficiency. Four rounds of data for a cohort of students are used to detect the long term math achievement gap between English Language Learners (ELLs), Former ELLs and NonELLs. Model-building process suggests that language demand in tests may have contributed to the big achievement gap between ELL and Non-ELLs. Long term and differential effects of other background variables are also detected.

Implication of the results and limitations of the technique are discussed.

DIFFERENTIAL LANGUAGE INFLUENCE ON MATH ACHIEVEMENT

by

Fang Chen

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2010

Approved by

____
Committee Chair

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ___________________________________

Committee Members ___________________________________

____

____

____

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to acknowledge several persons who have contribute to this dissertation.

Dr. Chalhoub-Deville is my advisor and has given me support throughout my Ph.D pursuit. Her language testing courses opened a door to a new world and helped shape my research interest. It is through her that I learned to apply all the research methodologies to practical issues such as language testing. Her broad knowledge and thorough understanding of the field inspired me to join and contribute. As the only female member on my committee, she also models a successful scholar who grows professionally and keeps a female's unique charm and perspective.

Dr. Ackerman's name is big in the field yet he is very humble and personable. His office is next door which makes it convenient for me to visit. Although he is extremely busy with all types of responsibilities, he always has the time to listen to my questions. I feel lucky that I have a great "neighbor" who is knowledgeable and ready to provide guidance whenever I stumbled during this project.

Dr. Luecht is the most challenging professor I have met who forces me to think beyond the books. He is very well-rounded in his knowledge and skills, and I enjoyed the many wonderful moments of "Aha" in his classes. He encourages me to explore, to think thorough, to try-and-fail and to be practical. He pushes me to the limits but never denies my effort. He inspires me to keep learning and growing.

I am very grateful that I took all the fundamental courses from Dr. Willse. His attention to detail helped me build a solid foundation that benefited me throughout this process. His R programming course equipped me with a new tool that makes this dissertation much better. His questions make me think and help me to clarify my thoughts.

Dr. Faldowski is from an outside department yet his knowledge on educational research and measurement issues is amazing. Whenever I ask him questions, I usually get an immediate response together with rich reference books or articles. He is an expert in regression and is a key force in my dissertation progress. I am very glad I took two courses from him, focusing on regression and longitudinal studies. As can be seen, they are closely related to my dissertation.

I would also like to extend my gratitude to Dr. Henson. He is not on my committee but I bothered him as much as if he were. He has also given me many valuable suggestions and reliable support. Other persons include my classmate, Devdass and Carolyn. Devdass introduced me to Dr. Faldowski's class, which then extended my interest to quantile regression for this dissertation. Carolyn volunteered to read my drafts and helped to make it precise.

Finally, I would like to thank my families for support. No doubt, my whole doctoral pursuit, including this dissertation, has distracted my attention from them. Without their understanding and support, I cannot achieve what I have today.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The first sentence in Chapter 9 of the most recent edition of *Standards for Educational and Psychological Testing* (hereinafter, “the Standards”, 1999) clearly states:

> *For all test takers, any test that employs language is, in part, a measure of their language skills. (p.91)*

This statement is true because all assessments employ language to measure student achievement. Students need the appropriate language skills to read the tests and sometimes to respond to open-ended questions. Consequently, all the test scores include variance introduced by the various level of language ability of students. This language ability is not the construct under examination in content area assessments but is confounded with students’ performances on these tests. Construct-irrelevant variance has attracted attention from many scholars who called for improvement of the psychometric quality of tests (Haladyna & Downing, 2004). In order to improve the content area tests, it is necessary to understand how language impacts academic performance in these assessments. The current study will focus on the relationship between language and math, but the method can be applied easily to other subjects to detect language impact in other assessments. The relationship between language and math has been found to be shifting rather than stable, although the direction of the shift was not clear. With the power of quantile regression, the differential language impact on math achievement will be fully explored. Longitudinal data will also shed light onto the long term language effect on academic performance that has not been well represented in past research.

Statement of the Problem

As Abedi and Lord (2001) concluded, “the interaction between language and mathematic achievement is real.”(p.232) This interaction is real for all grades, for both genders and for various ethnic groups. Overall, language ability is positively related to math achievement. One of the most heated discussion of this relationship is reflected through the achievement gaps between native English speakers (Non-ELLs) and English language learners (ELLs) in the K-12 grades. Literature has provided consistent evidence that ELL students scored lower than Non-ELL students in math assessments. This gap is regarded to be closely related to the limited language proficiency of ELLs (Abedi and Lord, 2001; Kato et al., 2004; Kieffer et al., 2009; Stevens et al., 2000; Wright and Li, 2008).

The math achievement gap between these two groups may be attributed to two parts: as the result of learning and the result of assessing. As the result of learning, students actually differ in math achievement because they did not learn the math content effectively; as the result of assessing, students are not so different in math achievement, but the tests underestimate it for some students because the language requirement in the test is too high for them to perform well (Bailey, 2000). There is no way to remove the first cause through test scores but the second possibility can be controlled with statistical techniques. This research analyzes the test scores as a source of information and tackles the assessment issue rather than the learning issue. The learning process is better studied through other means like think-alouds, observations, questionnaires or interviews.

Literature not only shows the gap between ELL and Non-ELLs, but also provides conflicting results on the changing status of the gap. The math achievement gap is found to be increasing in some studies yet decreasing in other studies as students move from lower grades to higher grades. There are several speculations on this discrepancy. For example, the ELL groups were defined according to different criterion in these studies. Not only did the number of categories differ, so did the sample in each study. The math measurements involved in each study usually differed from each other, which implies inequality of test specification, psychometric quality and language requirement. Other background factors may have also interfere with the relationship between language and math. All these differences naturally led to different results. Above all the differences, the relationship between language and math may not be static but instead change across grades as well as within grade. In addition, few studies directly controlled for the language proficiency of students, which means the impact of language in math assessment for individual students was ignored, resulting in an inaccurate representation of overall math achievement gap.

To counter-balance these issues, a better research design and methodology need to be used. For example, a cohort of students traced and measured by the same instruments for several years addresses the sample difference. A quantile regression methodology that models more than the central tendency can handle the differential language effect at various math achievement levels. Adding the language proficiency covariate solves the last limitation in past research.

Purpose of the Study and Research Questions

Deeper understanding between language and math facilitates critical decisions. For example, language may be found to have a bigger impact on math achievement of all students with low math ability. Therefore, additional language support should be given to all students regardless if they are ELLs or not. A longitudinal examination of the relationship may reveal that a math score with language influence directly controlled produces a different magnitude of math achievement gap than if the language impact is not controlled. This may then suggest a different way to describe students' math achievement and progress. Because test scores are frequently used to measure school effectiveness as well, these findings will also inform school level accountability decisions.

No matter for what purpose the test is designed, if the scores are assumed to reflect math achievement, then analysis should start with a purer or adjusted math component. For this purpose, it is suggested that the language variance be partialled out of the math score before the math achievement and growth can be properly described. Whether to study the relationship between the two variables or controlling one to report the other, regression is the natural choice of statistical method. Quantile regression is used instead of traditional mean regression for the many advantages to be dicussed later.

This study aims to answer the following three research questions:

1. How does language proficiency affect math achievement within and across grades?
2. How does math performance vary with respect to other background variables such as gender and socioeconomic status after language proficiency is controlled?
3. Does the math achievement gap between ELLs, former ELLs, and Non-ELLs increase or decrease as students move to higher grades?

Theoretical Background

English Language Learners and the Inequality in Education

"English language learners (ELLs)" is only one of the many terms used to refer to a specific group of students. This group is actually heterogeneous in first language, cultural background, family history, social economic status and educational orientations (LaCelle-Peterson & Rivera, 1994). They are also defined differently in different states (Goh, 2004) and according to different performance standards (Abedi, 2007; Chalhoub-Deville & Deville, 2008). Despite all these differences, these students share the same fact that they are still in the process of learning the English language and may have more challenges in academic achievement due to their limited language proficiency. The current research used the term "English language learners (ELLs)" rather than "Limited English Proficiency students" (NCLB, 2001) to eliminate any negative connotations of a deficit (Kieffer et al., 2009).

The high educational risk for ELLs has been noticed and documented by different sources: ELLs have high risk of academic failure and school dropout (Garcia, 2000); ELLs score lower than main-stream students on national assessments in reading, math, and science (Kieffer et al., 2009; US. Department of Commerce Bureau of Census, 1993); proportionally more ELLs are receiving special services (Kretschmer, 1991); and ELLs have lower rates of college entry and progress at the university level (Astin, 1982).

Studies have been conducted to try to trace the cause of these phenomena. Some insights have also been provided. For example, scholars believe many critical decisions concerning ELLs focus exclusively on test scores but the reliability and validity of standardized test scores for ELLs are problematic (Abedi 2002; Chalhoub-Deville & Deville, 2008; Lam, 1993). As a result, ELLs' achievements may be underestimated. Others found that ELLs lack the opportunity to learn the content knowledge (Herman & Abedi, 2004; Wright & Li, 2008) although they might have achieved appropriate language proficiency by the time they need to go to college (Pennock-Roman, 1990). The actual inequality in education is the result of both learning and assessment. The common cause of these, however, is the impact of language demands in academic settings (Abedi and Lord, 2001; Kato et al., 2004; Matinez et al., 2009; Stevens, Butler, & Castellon-Wellington, 2000; Wright and Li, 2008).

Language Proficiency in Academic Settings

Whenever language proficiency is discussed, BICS and CALP are the two terms that have to be distinguished. Cummins first named them and emphasized the difference on several occasions (1979a; 1999). BICS refers to Basic Interpersonal Communicative

Skills and CALP refers to Cognitive Academic Language Proficiency. The former deals with everyday social interaction and the latter relates more closely to classroom activities. BICS can be learned rather quickly within two years at peer-appropriate level but CALP takes a much longer time for immigrant children (Collier, 1987; Cummins, 1981a, 1981b, 1984; LaCelle-Peterson & Rivera, 1994).

Cummins (1999) pointed out that the two concepts are not mutually exclusive. The distinction is made to emphasize the different patterns of development. While BICS such as phonological skills and fluency may reach a plateau quickly, CALP skills such as literacy and vocabulary continue to grow throughout schooling. In this sense and in the context of educational settings, language proficiency leans heavily towards CALP. The terminology was updated to Academic Language Proficiency (ALP) in current ELL literature and is sometimes referred to as Academic English since English is the language of instruction in the U.S. system. ALP is now widely accepted to be the key to school success because it is required to understand teacher talk, participate in class and handle content assessment (Bailey & Butler, 2003; Stevens, Butler, & Castellon-Wellington, 2000; Wilkinson & Silliman, 2000). ALP includes all four language domains, namely listening, speaking, reading and writing. Reading and writing are more important than listening and speaking in common assessment settings because they are the skills usually necessary for students to understand the questions and to respond to them.

Theory on language proficiency has developed beyond the simple and rigid way of decomposing it as Cummins did. Other concepts such as communicative language competency (Bachman, 1990; Bachman & Palmer, 1996) and language-in-use-in-context (Chalhoub-Deville & Deville, 2006) introduced a framework of a changing construct depending on the environment rather than a static one. On the other hand, definition of ALP is urged to be more specifically related to academic contents (Bailey, 2000; Bailey & Butler, 2003). All these reveal the challenge in developing an assessment of ALP. The discussion on this topic will not be replicated here but it is important to be aware of because it limits the choice of the key independent variable for this research.

Ideally, to control for academic English proficiency in academic assessment, a high quality ALP test score should be used. However, this is not widely available in practice. The current research for example, uses a reading score as a proxy of ALP. This is a reasonable practice for three solid reasons. First, there are few high quality ALP tests available. Research on academic language proficiency has shown the challenges in creating such instruments (Bailey, Butler, & Sato, 2005; Chalhoub-Deville & Deville, 2008). Lacking guidance is just one of them, which explains the failure of existing tests to meet desirable psychometric quality (Abedi, 2002, 2007). Second, use of ALP assessment seems to be limited to the ELL population only. However, language affects math achievement for everyone (Abedi & Lord, 2001; Freeman & Crawford, 2008). If language proficiency is to be studied, it is better to have the measurement for everyone including Non-ELLs. Reading assessment is usually conducted on everyone thus can serve this purpose. In addition to this, the heterogeneous difference within ELLs can also be taken into consideration by the direct measure of language proficiency at the individual level rather than a rough group membership of being ELLs or Non-ELLs

(Chalhoub-Deville & Deville, 2008). Third, understanding written text is the first important form of language proficiency for cognitive functioning (Mestre, 1988). Reading is the skill that is inseparable for performance in tests while other modes such as speaking or writing are usually not as involved or critical for math assessment. In this sense, reading is regarded as a close substitute of academic language proficiency and does not overestimate language influence in assessment by involving irrelevant language domain skills (listening, speaking and writing).

Accommodation for ELLs

Standardized tests have been widely accepted in educational assessments because they can increase reliability and reduce random measurement error due to testing procedures. The key features are the standardization of test form, test administration procedure and predefined scoring rubrics (Goh, 2004). However, standardized tests have been shown to be inappropriate for ELL population for several reasons. For example, ELLs may not be represented in the norming population (Davison, 1994; Stevens etal, 2000) and the meaning of standardized tests scores may not be the same for ELL versus Non-ELL students (LaCelle-Peterson & Rivera, 1994). Also, assumptions about students who take standardized tests are obviously violated for ELLs (Lam, 1993). One of the assumptions is that test takers have no linguistic barriers that inhibit their performance on the test. This assumption was rarely supported by standardized tests. On the contrary, because of the confounding issue of language in content knowledge assessment, it cannot be judged whether student performance on standardized tests reflects their language ability or content knowledge (Abedi et al., 2005; Bailey & Butler, 2003; Kieffer et al., 2009; Rivera, et al., 2006).

NCLB and the *Standards* (AERA, et al, 1999) both support the necessary accommodation for ELLs in standardized testing to accurately measure their achievements and progress. In the *Standards* (AERA, et al, 1999), accommodation is defined as:

> *the general term for any action taken in response to a determination that an individual's disability requires a departure from the established testing protocol. (p101).*

The disability here is the limited language proficiency for ELLs. Goh (2004) summarized four possible accommodations for ELLs, including setting modifications, timing and scheduling modifications, presentation modifications and response modifications. Certainly, all these accommodations assume that that language proficiency limitation can be easily overcome through some procedural help during testing. However, research has shown that only *linguistic* accommodations made a difference in student performance while other common practices such as extra time did not help (Abedi, 1999a, 1999b; Abedi & Hejri, 2004; Francis et al., 2006; Menken, 2000). As already mentioned, the gap between the academic achievement of ELLs and Non-ELLs can be traced to both the learning process and the assessment process. Accommodation just in the testing procedure is not enough to eliminate irrelevant factors in assessment.

Consistent with this insight, new programs have been started to reduce the linguistic burden for students both during learning and assessment. Help with English

Language Proficiency (HELP) Math program (Freeman & Crawford, 2008; Tran, 2005) is a Web-based curriculum aimed to provide interactive lessons and “essentialize mathematical vocabulary and academic concepts so that students can better understand the content” (Freeman & Crawford, 2008, p.5). Other programs such as Obtaining Necessary Parity through Academic Rigor (ONPAR) aims to use innovative computer-based items with minimal language requirement to assess ELLs (Kopriva et al., 2009). However, to ensure validity of test score interpretation, test format should be consistent with the teaching format. That is, the way the students are assessed should be the same as they are taught. Both HELP and ONPAR are valuable researches but they are not used in practice on the same students. Unless the students are taught and assessed with the same kind of support, the validity question of what a test is measuring remains a challenge.

Researchers have pointed out that all accommodations require extra resources and money (Abedi, Hofstetter & Lord, 2004; Abedi & Lord, 2001). The above mentioned innovative instruction and corresponding assessment are not widely used in the U.S. Cost may be one of the reasons. Before all students have access to these types of innovative instruction and assessment, another feasible approach is needed to better describe students’ achievement with efficiency.

The approach recommended in this research is to control for the language proficiency of students and report residual of content test scores (math in this case) after the language proficiency is partialled out. In this way, students’ achievement can be depicted independent of their language ability. Whether students’ achievement is due to the learning or the assessment is not the topic here. The interest here is to describe the math achievement and do it in a more accurate way. This approach is named “accommodation in score reporting” and can be regarded as an alternative to accommodation for ELLs. To follow the principal of fairness (AERA, et al, 1999), this accommodation in the form of partialling out the language impact before reporting math should be done for both ELLs and Non-ELLs since literature has shown that language affects math for all students (Abedi & Lord, 2001; Freeman & Crawford, 2008; Kiplinger, Haug & Abedi, 2000).

Summary

Language proficiency affects content knowledge learning and assessment especially for ELLs. Regular standardized tests failed to take students’ language proficiency into consideration. Test results may not reflect students’ achievement accurately. In addition to this, the relationship between language and math may vary for students within and across grades. To better explore the possible differential influence of language on academic achievement, new research design and analysis technique will benefit. Reading scores can be used as a reasonable proxy of academic language proficiency for all individuals. When reporting content test scores, reading can be controlled for everyone to generate a more accurate description of content area achievement independent of language. Accommodation in score reporting is an alternative to traditional accommodation for ELLs. It may serve as a viable tool, before innovative instruction and assessments are both in place in practice, to eliminate construct-irrelevant variance in assessments due to the language requirement.

Assumptions of the Study

1. The math and language items are assumed to measure the construct of math and language proficiency and no other factors.
2. There is no measurement error.
3. Variables are measured independently from each other.

Limitations of the Study

The limitations of the study are:

1. All students are assumed to have the opportunity to learn. Once the language requirement in math scores are partialled out, the residuals reflect students' actual math achievement. In reality, students with limited language proficiency may be doubly punished by the language disadvantage through both learning and testing.
2. Reading score is used as an approximate indicator of academic English proficiency. The best language measure for the purpose should be on the academic language proficiency and specific to the math subject.
3. The students in this study are a cohort followed for eight years since early ages. The achievement gap between ELL and Non-ELL groups might be different from what is observed in any current grade in the U.S. The meaning and generalization of the specific achievement gap should be interpreted with caution.

Chapter II

LITERATURE REVIEW

Research on the Relationship between Language and Math Achievement

A Confirmed Relationship

Research has provided all types of evidence that language proficiency affects math achievement. This has been documented and confirmed for all grades including 1st (U.S. Department of Education, DOE, 2008), 2nd (Abedi etal, 2005), 3rd (Cottrell, 1968; Brown, 2005; Butler & Castellon-Wellington, 2005), 4th (Chang, Singh & Filer, 2009; Fry, 2007), 5th (Chang et al., 2009), 6th (Balow, 1964; Freeman & Crawford, 2008), 7th (Freeman & Crawford, 2008), 8th (Abedi & Lord, 2001; Fry, 2007), 9th (Abedi et al., 2005), 10th (Abedi, Leon & Mirocha, 2003) and 11th grade (Butler & Castellon-Wellington, 2005).

Impact of language has also been found for different populations including ELLs and Non-ELLs, boys and girls and both students with high and low social economic background (Abedi & Lord, 2001; Kiplinger et al., 2000). There are also researches of its impact on specific immigrant students such as Cambodian (Wright & Li, 2008), Hispanic or Spanish (Freeman & Crawford, 2008) and Asian students (U.S. DOE, 2008).

Through a representative sample of studies, Aiken (1971;1972) summarized the correlation between language and math achievement which ranged from .40 to .86.

Secada's (1992) summary showed the correlation ranged from .20 to .50. All the correlations were positive and statistically significant. The difference in the correlation coefficients across all studies was due to the various measures of general and specific reading abilities (Aiken, 1972). As to be elaborated later, this can also be due to the different grades involved. In all, these studies confirmed that there is a non-negligible relationship between language and math achievement.

In summary, it is seen that language ability affects math achievement for all population groups at various grades. This issue is quite prevalent in the ELL community due to the clear language disadvantage of ELL students.

A Changing Relationship

Although there is a great deal of studies detecting the relationship between language and math, many more issues need to be explored to better understand *how* language affects math achievement. One issue is that the relationship seems to be changing rather than static.

Ausubel and Robinson (1969) pointed out that at least at early stages, kinesthetic images serve as the base of math learning for understanding arithmetical ideas in particular and the inductive process of concept formation in general. Naturally, it is reasonable to assume that language ability might not affect math learning that much at early math learning stages. However, as students move on to higher stages in math learning, concepts may not be easily visualized and more complicated reading skills become more and more vital. If this is true, newly classified ELLs at higher grades might have more difficulty with math than the newly classified ELLs at lower grades. This partly explains the gap between ELL and Non-ELL students in math achievement between, for example, 4th grade versus 8th grade (Fry, 2007). In that study, the math achievement was smaller at 4th grade than at the 8th grade.

Interestingly, some studies found the group-based math achievement gap increased at higher grades (Abedi et al., 2005; Butler & Castellon-Wellington, 2005; Fry, 2007) while others found the gap decreased over time (Chang et al., 2009; Galindo 2009; Han, 2008). This discrepancy in findings might be due to several confounding issues.

First, ELL status was defined differently in these studies. For example, Abedi et al (2005) divided students into four categories: Students with disorders (SD), Limited English Proficiency students (LEP), students who are LEP and SD and None of the above. Chang et al (2009) divided their students into three groups: English-only (English), Dual-language-speaking (DUAL) and English-Language-Learner (ELL). These different criteria of classification obviously affected the meaning of gap between ELLs and Non-ELLs.

Second, the math measures involved in these studies were different. Abedi et al (2005) conducted their research at two sites. The math instruments used were Iowa Tests of Basic Skills (ITBS) at one site and Stanford Achievement Test Series, Ninth Edition (Stanford 9) at the other. Butler and Castellon-Wellington (2005) also used Stanford 9. Fry (2007) used the National Assessment of Educational Progress (NAEP) math assessment and Han (2008), Chang et al (2009) and Galindo (2009) used the instrument created for Early Childhood Longitudinal Study- Kindergarten Cohort (ECLS-K). All these instruments might differ in psychometric quality. In terms of linguistic features, some of them might put ELLs at a disadvantage more than the others. As a result, the raw math achievement gaps are not comparable between studies.

Third, ELL composition in each study is different. For example, Abedi et al (2005) studied the math gap for students from 2nd to the 11th grade. However, every grade was formed by a separate group of students that belonged to that grade at the time of assessment. Fry (2007) found out that the decline in achievement from elementary to middle school was partly due to change in the composition of samples. He noticed that many former ELL students who caught up were no longer categorized as ELLs as they moved on to higher grades. At the same time, newly arrived immigrants were added to the ELL groups. In other words, the gap in math scores in these studies may not reflect the *achievement* difference but the difference in the *initial status*. When the gaps are then compared across grades, it is hard to say whether the achievement gap is really increasing or decreasing. In comparison, Han (2008), Chang et al (2009) and Galindo (2009), used a longitudinal data where the students were traced for several years. The difference between these students might more accurately reflect the achievement gap. These latter three studies imply the necessity of longitudinal data to better capture a long-term relationship because they may reduce interference from external factors such as a changing demographic group.

Fourth, none of the studies used a continuous measure of language ability as a covariate while estimating the math achievement gap. It is reasonable to expect that overall, the gap in math scores will be less if the construct-irrelevant variance due to language is removed.

If Ausubel and Robinson (1969) were right, challenges in learning the math content will be more difficult to overcome for newly arrived ELLs at higher grades than at lower grades. This phenomenon will not be detected directly in this research since the same students were traced and only test scores are analyzed. However, reading and math are separate skills in the end. Once a certain language proficiency threshold is reached, the influence of language on math may start to decrease. The predictive power of language on math scores may decrease or the achievement gap will shrink once language proficiency is controlled. Either way, disappearance of the gap as reflected by scores is not likely to happen statistically due to the large sample size usually found in these type of studies. The question is: if the same students are traced and the language impact in assessment removed, does the math achievement gap diverge or converge?

Compared to the amount of attention on the math achievement gap across grades, much less was directed to the other aspect of a changing relationship that is happening *within* a grade. Freeman and Crawford (2008) demonstrated that a revised math curriculum that removed linguistic and cultural barriers was more effective with ELLs with higher language proficiency than those with lower language proficiency. Kopriva et al (2009) reported that math items where sentences reduced to phrases actually added to the difficulty of the items compared to items where language were modified but still in a sentence format. More research needs to be done to see how the language impact differs at various language and math levels within a grade as well as between grades. This research is expected to fill in this gap in literature which is possible with the quantile regression technique.

Other Variables of Interest

Secada (1992) reviewed various data on the math achievement gap and indicated that language proficiency does not explain the difference between students completely. Several other factors consistently appear in literature, which seem to contribute to math score variance above and beyond language or ELL status. These factors include socioeconomic status (SES), parental educational level, gender and race-ethnicity. An understanding of these factors helps with the decision to include or exclude them for the current research purpose.

There is rich evidence that SES affects students' math achievement (Abedi & Lord, 2001; Abedi et al., 2005; Brown, 2005; Butler & Castellon-Wellington, 2005; Chang etal, 2009; Liu & Wilson, 2009; Tate, 1997; Wright & Li, 2008). Some research even found a differential effect of SES. For example, Brown (2005) used a multiple regression analysis to predict math achievement for ELL and non-ELL students. He found out that low SES students performed similarly whether they were ELLs or not, but for high SES students, Non-ELLs outperformed ELLs even after language scores were controlled. Literature on SES is quite consistent on its impact, which highlights the necessity of its inclusion in any educational research. However, SES in many of these studies was only a categorical variable, dividing students into either high or low SES groups (Abedi & Lord, 2001; Abedi et al., 2005; Butler & Castellon-Wellington, 2005). A continuous SES regressor as in the current research might reveal more information of its impact.

Parental educational level is a strong predictor of students' math achievement, even stronger than SES according to Abedi et al (2005). A brief from the Department of Education (2008) showed specifically that ELLs whose mother had a bachelor's degree or higher made greater gains than their peers whose mother had a lesser degree. Parental educational level is usually available as an ordinal variable but frequently dichotomized to detect the *desired* group difference (e.g. Abedi, et al 2005). In the current research, the SES index is a continuous composite including parents' income, educational levels and occupations. Preliminary analysis also showed no statistical significance of parental educational level once SES is in the model. Based on these facts, parental educational level is removed from the current research. Its influence is assumed to be subsumed under SES.

Gender is a variable frequently studied, but conclusions on gender differences tend to be controversial. There are research results favoring boys in math performance (Benbow & Stanley, 1980; Gallagher et al., 2000; Leahey & Guo, 2001; Mau & Lynn, 2000), or favoring girls (Ginsburg & Russell, 1981; Kaplan & Weisberg, 1987). At the same time, there is no shortage of research showing no gender difference at all (Geary, 1994; Lummis & Stevenson, 1990; Tate, 1997). Recent studies also demonstrate a varying impact of gender on math achievement dependent on the math proficiency level of students within or across grades (Benbow, 1992; Hyde, Fennema & Lamon, 1990; Leahey & Guo, 2001). Just like the shifting relationship between math and reading, this discrepancy of gender impact on math achievement may be related to the different samples and or instruments used. Similarly, language impact was not controlled in these studies which might have affected math score comparisons between genders. Similar to the issues between language and math, quantile regression with longitudinal data may reveal more about gender effect in the current research.

Race or ethnicity comparisons are usually limited to a few groups, perhaps because of the relative small size of other populations. For example, math achievement gaps are well documented but usually between White, African-American, Hispanic and Asian groups (Abedi et al 2005; Haile & Nguyen, 2008; Tate, 1997). Although there are numerous ways to break down the groups, this research will follow the tradition to have mainly four race-ethnicity groups. Sample size limited other groups to be a meaningful separate unit. However, although Hawaiians and other Pacific Islanders, American Indians and Alaskans do not meet the common definition of ELLs, their ethnicity may have put them at disadvantage due to geographic isolation from main stream Americans. For the current research, they were coded as a separate group to represent the maybe-disadvantaged native Americans.

There are other possible variables that may also affect the math achievement differently. However, following the principle of parsimony and to focus on the current research question of interest, discussion of a possible list will stop here. The factors to be considered in this research are regarded as comprehensive enough without hindering the interpretability of results. To conclude, language proficiency, SES, gender and race-ethnicity are kept as independent variables in the current research. Parental educational level is embedded in the SES index and not represented by itself. Descriptions of the selected variables and measures will be provided in detail in the next chapter.

Limitations of Past Research

As explained before, there are many reasons for the conflicting results on the influence of language on math, especially in the long run. Limitation in research design and analysis techniques are the main source. Because issues like samples and variables included or excluded in the research affect both the soundness of the research design and the specific models used, they are discussed under both sections of limitations in research design and limitations in statistical models.

Limitations in Research Design

Most of the research design limitations have already been discussed in detail when reviewing the conflicting results on the diverging versus converging math achievement gap. For example, the sample difference as well as instrument difference contributed to the inconsistency between studies. These differences contaminated conclusions not only for achievement gaps but also for background gender effects. Several key issues are revisited below. The solution to these issues is the longitudinal data and a direct measure of language proficiency.

First, as the focus of many studies, the relationship between language and math was revealed only in a roundabout way without a *direct* measure of language ability. Many research results concluded on the impact of language just with the average difference at the group level. For example, Abedi and Lord (2001) modified some items from the 1992 NAEP math assessment that are identified to be potentially problematic for ELL students. The linguistic complexity of some items was modified. Students' score differences between the original and the modified items were assumed to represent the language influence on math. This depiction of the language influence is only a rough sketch of the more complex picture. A better way is to use a continuous language proficiency measure so that a wide range of language proficiency can be studied (Chalhoub-Deville & Deville, 2008). This research uses the reading score as a proxy at the individual level. It benefits not only as a direct measure of language proficiency but also a measure with richer and finer information along a whole distribution.

Second, most studies are cross-sectional and focused on one grade (Abedi et al., 2005; Brown, 2005; Butler & Castellon-Wellington, 2005; Fry, 2007). Different samples were used in different grades and sometimes different variables were considered at each grade. The impact of language on math cannot be compared across grades with confidence. In the same vein, changes in achievement gaps across grades might not reflect changes of achievement but of the status quo for each new grade because of demographic change. For example, Fry's (2007) study involved students from grade 2 to 11 but they were different students at each grade just measured at the same time. The composition of samples varied in percentage of ELLs. Results based on this type of cross-sectional studies cannot provide convincing evidence of a closing or diverging achievement gap between ELLs and Non-ELLs. Longitudinal studies have the advantage that the sample stays the same. Many background factors related to students' academic achievement can be assumed to be quite stable for the same students and the math achievement can be interpreted with more confidence to reflect the progress in math. For this reason, the Early Childhood Longitudinal Studies- Kindergarten Cohort (ECLS-K) is selected for the current research.

Third, there were studies in the past that used longitudinal data but the interest was on the math achievement regardless of language proficiency variability among individual students (Chang etal, 2009; Galindo 2009; Han, 2008). Past results confirmed the influence of language on achievement gaps and the shifting language impact, yet these two phenomena were never explored within the same model. The ECLS-K dataset has the reading scores as well as the math scores at every assessment time point. By including the reading scores when modeling math, a long term language impact can be observed. In addition to that, since language proficiency affects student performance in standardized math assessment, the math growth based on the unadjusted score is not as accurate as a score where language proficiency is controlled. By including the reading score as a covariate, a purer math trajectory will result[1].

Limitation in Statistical Modeling

In order to detect the shifting language influence on math achievements, various methods were tried in the past to solve the puzzle from different perspectives. For example, Abedi and others (2005) used simple regression, multiple regressions, principle component analyses and canonical correlational analyses on the same dataset to try to see how language affects math differently at different grades. Coordinating all results from various methods was already a challenging task. In addition to that, sometimes each

[1] Note that the same regression models serve two purposes: the slope reveals the relationship and the residual represents the pure math achievement after accounting for language proficiency.

analysis used only part of the variables. The fact that all the variables are not studied at the same time can lead to conflicting conclusions. This latter point was very well reflected in Abedi and Lord (2001). Abedi and Lord (2001) conducted two separate 2-way ANOVAs. One was to see the effects of ELL status and SES status on math and another to see ELL status and item types (language factor) on math. The first analysis produced a significant interaction term which prevents meaningful interpretation of main ELL effect but the second one found a consistent ELL effect. This type of discrepancy about ELL effect is hard to reconcile between separate analyses. A better way is to include all variables in the same model and analyze them simultaneously.

The biggest statistical problem with past research, however, was that they all revolved around a mean or an average pattern. Sometimes groups were compared just on mean math scores (Abedi & Lord, 2001; DOE, 2008; Fry, 2007); sometimes mean regressions were used (Abedi, et al., 2005; Brown, 2005; Butler & Castellon-Wellington, 2005; Chang et al., 2009; Han, 2008). In all, an average pattern was used to represent a whole distribution of possible patterns. This is fine when conditional distributions (or the error distribution) are normal. However, if there are changes in higher order moments such as skewness or kurtosis of the distribution, the median may be a more appropriate measure of central tendency than the mean (Edgeworth, 1888; Fox, 1997; Hao & Naiman, 2007; Koenker, 2005). Some studies have already hinted that language does not affect low math ability students the same as it affects high math ability students (Abedi et al., 2005; Ausubel & Robinson, 1969; Butler & Castellon-Wellington, 2005; Chang et al., 2009; Fry, 2007; Galindo 2009; Han, 2008). If this speculation is true, locations other than the median are also of interest. A natural thought is to truncate the population into subgroups based on the unconditional math scores and conduct several mean regressions. This approach, however, could create biased parameter estimates as thoroughly argued by Heckman (1978). When a differential effect (e.g. language) rather than a constant is expected for the independent variable, a quantile regression (QR) is more appropriate than a mean regression (MR) (Hao & Naiman, 2007; Koenker, 2005; Koenker & Hallock, 2001).

MR requires several assumptions which in reality are not always met. The normal distribution of errors and homoscedasticity are two of them. Violation of assumptions can produce misleading results and sometimes even prevent programs from running. For example, Butler & Castellon-Wellington (2005) reported that a MANOVA on Grade 3 data did not run because of the coexistence of unequal variances and unequal sample sizes between groups. Violation of homoscedasticity was believed to be the issue. If the normality assumption holds, violation of homoscedasticity may be accommodated through multilevel regression models. If normality does not hold, ordinary least square (OLS) in mean regression still produces unbiased and consistent estimates but is not the most *efficient* estimator (Fox, 1997). In that case, again, a robust regression method such as quantile regression is an alternative (Hao & Naiman, 2007; Koenker, 2005; Koenker & Hallock, 2001).

In brief, technical limitations in past research can be overcome by quantile regression and by including all the variables of interest in the same models. Due to the complexity of quantile regression as a new approach, a separate section is devoted to introduce it in more detail. The goal is to provide a basic understanding of quantile regression and its utility through some examples.

Quantile Regression: Basics and Applications

Quantile is an equivalent term to percentile (also called fractile) where the median is the 50th quantile. Similarly, 25th and 75th quantiles correspond to the first and third quartiles. Quantile regression modeling (QRM) is a term for a series of quantile regression alternatives. Roger Koenker (2005), the author of the first book devoted to quantile regression, traced quantile regression back to Boscovich, even prior to the discovery of least squares commonly used for mean regression. In that first attempt to "ever *do* regression" (Koenker, 2005, p.2), Boscovich estimated the slope coefficient through a process which Laplace later noted as the computation of a *weighted median*. However, the model was an interesting mixture because while the slope was estimated based on the median, the intercept was still estimated as a mean. In 1888, Edgeworth improved Boscovich and Laplace's idea by proposing a process to minimize the sum of absolute residuals in both intercept and slope parameters (Koenker, 2005). The first complete quantile regression thus started.

Clearly, quantiles are order-statistics and are more resistant to outliers. If errors follow a Gaussian (normal) distribution, results of MR and QR at the median coincide. If errors are not normally distributed or homoscedastity does not hold, QR provides a more efficient estimate. In addition to that, QR can detect the differential effects of independent variables on the dependent variables that MR cannot detect. This quality of

QR enabled in-depth research in many fields. For example, Koenker and Hallock (2001) studied the determinants of infant birthweight. MR revealed that baby boys weighed more than 100 grams than girls on average. However, although this direction of disparity was consistent across the weight distribution, QR revealed that the magnitude of the disparity was less at the 5th quantile (45 grams) than at the 95th quantile (135 grams).

QR has helped to advance knowledge in many fields. It is a frequent tool in economics and it is regarded as "the standard tool in wage and income studies in labor economics" (Yu et al., 2003, p.339). For example, Koenker and Bilias (2001) found the Bonus System in Pennsylvania can shorten the protected unemployment period for some of the unemployed but not all. Chevapatrakul and others (2009) used quantile regression to confirm the Taylor Principle in finance. In medicine, QR has been used for studying gender and demographic covariates effects for end-tail quantiles of the population (Abreyeya, 2001; Austin et al., 2005). It is also used for developing medical reference charts (Cole, 1988) and growth charts (Wei, et al., 2006). QR has also been applied to environmental studies (Pandey & Nguyen, 1999) and to survival analysis (Koenker and Geling, 2001; Yang, 1999). To the author's awareness, however, application of QR in the educational measurement field is rare. QR has never been used for language-related research or assessment issues. As has been established, QR is a natural tool to study the shifting relationship between language and math (or the differential effect of language proficiency on math). It is hoped that this research will fill the gap in educational research for QR applications.

To fully appreciate the advantage of QR for the current research, the following sections are devoted to introduce basics of QR modeling (QRM)[2], balanced between details and brevity. When appropriate, the variables for the current research are used to make the concept more concrete. Only one independent variable (x) is considered in the introduction but extension to more independent variables will be demonstrated in the actual research later. QR can be linear or non-linear, parametric or non-parametric. This introduction, however, will be limited to linear QR models. The model here is regarded as semi-parametric where the deterministic portion (prediction) still assumes a parametric form although the error term does not (Cade & Noon, 2003).

QRM in Equations

Using one covariate as an example, a simple mean regression model can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (2.1)$$

All the data are used to find one regression line that minimizes the error term or the least squared distance (LSD) objective function. Algebraically, the goal is to find the point where the first derivative of the mean squared deviation is zero with respect to the mean. Graphically, the resulting regression line is a line that minimizes the sum of squared vertical distances of all response observations to the line. The best fitting line is actually the one that passes the expected means of the response distributions conditioned at every covariate value.

[2] The organization of this section follows the framework of Hao and Naiman (2005).

Analogically, quantile regression models can be written as

$$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)} \qquad (2.2)$$

The only notational difference between Equation 1 and 2 is the extra superscript '*p*', which specifies the *p*th quantile regression model.[3] Depending on the quantiles of choice, the QR regression lines are different. In practice, usually a whole set of quantile regression models are compared to detect the different covariate effect on the dependent variable at various quantiles of response distribution. Still, all the data are used for every quantile regression modeling[4]. The residuals are still minimized but by minimizing the *absolute* distance rather than the *squared* distance. The best fitting line is the one that passes the conditional *p*th percentiles of the response distribution. Taking the current research as an example where reading score is the independent variable and math the dependent variable, the best fitting line for *p*=.5 passes the conditional medians (50th percentile) of the math score distributions. In other words, half of the math scores lie above the line and half below the line. The same concept extends to other quantile regression models at other *p*s.

QRM Parameter Estimation

In mean regression modeling (MRM), estimates of the intercept and slope coefficients of the best-fitting line are the ones that minimize the sum of squared errors

[3] Koenker and other authors used τ rather than p. p is kept here for obvious meaning of "percentile".

[4] It is a misconception that only a subset of the observation is used for every quantile regression. All observations are used to locate a quantile since it is the pth value in the ordered observation. Also, the quantile regression analysis is a minimization of weighted sum of absolute residuals that involves all the observations.

$$\sum_i^n \varepsilon_i^2 = \sum_i^n (y_i - (\beta_0 + \beta_1 x_i))^2 \qquad (2.3)$$

The estimates can be shown to be $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \dfrac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$.

When linearity, constant variance and independence of *x* values are true, ordinary least square estimation coincides with maximum-likelihood estimation. The corresponding slope coefficients are regarded the best linear unbiased estimator (BLUE) of the population parameter.

In QRM, estimates of the intercept and slope coefficients corresponding to the best-fitting line are the ones that minimize the *weighted* sum of *absolute* errors

$$\sum_i^n w_p |\varepsilon_i| = \sum_i^n w_p \left| y_i - (\beta_0^{(p)} + \beta_1^{(p)} x_i) \right| \qquad (2.4)^5$$

$$\text{where } w_p = \begin{cases} p & when \quad y_i \geq (\beta_0^{(p)} + \beta_1^{(p)} x_i) \\ 1-p & when \quad y_i < (\beta_0^{(p)} + \beta_1^{(p)} x_i) \end{cases}$$

Or $\quad p \sum_{y_i \geq p} \left| y_i - (\beta_0^{(p)} + \beta_1^{(p)} x_i) \right| + (1-p) \sum_{y_i < p} \left| y_i - (\beta_0^{(p)} + \beta_1^{(p)} x_i) \right|$

When *p*=.5, this simplifies to $\sum_i^n \left| y_i - (\beta_0^{(0.5)} + \beta_1^{(0.5)} x_i) \right|$.

[5] Koenker's notation for this concept is $\sum_{i=1}^n \rho_\tau (y_i - \xi)$. Notation used here are consistent with what they commonly mean in social science field. This is decided to facilitate understanding and communication among general readers.

The solution that minimizes the weighted absolute distance is when $\hat{y}_i = \beta_0^{(p)} + \beta_1^{(p)} x_i$ equals the *p*th percentile. Further details can be found in Koenker (2005) and Hao and Naiman (2007).

Several algorithms are available to estimate the quantile regression parameters such as simplex (Koenker & d'Orey, 1987), interior point (Portnoy & Koenker, 1997), and smoothing method (Chen, 2007). The default algorithm in Quantreg in R and SAS are both simplex. However, this algorithm is computationally demanding and is not recommended for sample size larger than 5000 (SAS, 2008, p.5380). For larger sample size, interior point is faster.

QRM Standard Error Calculation and Confidence Interval

Once the coefficients are estimated, standard errors are calculated to test the statistical significance of the coefficient estimate $\hat{\beta}_1$. The intercept usually is not of interest in hypothesis testing and is not discussed here.

In MRM, the standard error for the coefficient β_1 is calculated by assuming the normal distribution of the error term. That is, the ε_i in equation 2.1 is regarded as independently and identically distributed (i.i.d.) across all covariate values. (In that sense, the subscript "*i*" can be dropped from the equation.) $S_{\hat{\beta}_1}$ is the estimated standard error of β_1 and the $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$ is assumed to follow a Student's t distribution with n-k degrees of freedom (k is the number of all coefficients plus an intercept). Consequently, the 100(1-α)% confidence interval for β_1 is calculated in the form $\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$.

The very motivation for the development of QRM is that the conditional response distribution is skewed. More importantly, the error terms do not follow i.i.d. Traditional approach for standard error thus is not appropriate and can be replaced by a bootstrap (Efron's, 1979) technique which does not necessarily require a specific form of distribution. The observed data set is regarded as the population. One method is to bootstrap pairs of observations (e.g. a reading score with a corresponding math score) from the observed data repeatedly and generate multiple samples. Every sample gives a parameter estimate. In this way, a distribution of the $\hat{\beta}_1 s$ can be collected. The standard deviation of these $\hat{\beta}_1 s$ is taken as the standard error of the parameter β_1 . The confidence interval can be calculated following a standard normal distribution in the form of $\hat{\beta}_1 \pm z_{\alpha/2} s_{\hat{\beta}_1}$. Another approach to determine the confidence interval is to take the empirical values from the same distribution of the estimated $\hat{\beta}_1 s$ and locate the corresponding empirical percentiles. For example, the 95% confidence interval of the parameter β_1 starts from the 2.5th percentile of the estimated $\hat{\beta}_1 s$ in the distribution and ends at the 97.5th percentile of the estimated $\hat{\beta}_1 s$.

The approach for SE described above is the standard xy-pair bootstrap. Within the bootstrap family, there are also other versions developed for QRM such as Parzen, Wei and Ying's (1994) version of the xy-pair bootstrap, the Markov chain marginal bootstrap

(MCMB) of He and Hu (2002) and Kocherginsky, He, and Mu (2005). Non-bootstrap methods are also developed that employ rank score function, sparsity function or kernel estimation of Huber sandwich (Powell, 1991). For more details about these methods, please refer to Koenker (2005).

Hao and Naiman (2007) gives a guideline that the number of bootstrap samples should be between 50 and 200 for standard deviation estimation and between 500 and 2,000 for a confidence interval. Fox (1997) recommended 100 to 200 for standard error and 1000 to 2000 for confidence interval. SAS uses the MCMB method for both SE and CI estimation but cautions its appropriateness only for large samples with at least 5000 observations and/or 20 variables.

Hypothesis Testing

For large sample size, after standard errors are calculated, hypothesis testing for the significance of a single covariate takes advantage of central limit theorem and follows the regular regression procedures. The *t*-statistics are calculated following

$$t = \frac{\hat{\beta}_1 - \beta_1^{null}}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \qquad (2.5)$$

and compared against the critical *t* with *n-2* degrees of freedom under the null distribution.

The bootstrapping method described above can produce a covariance matrix of the cross-quantile estimates. This matrix can be used for another type of hypothesis testing to see whether any difference between the slope coefficients for the same covariate is statistically different across quantiles. For example, does reading ability predict math achievement the same way (e.g., is the slope coefficient the same) for a student who is at the 10th percentile of the math score distribution versus a student who is at the 90th percentile of the math score? This is called test of equivalence in the current research. The Wald statistics is for this purpose.

$$\text{Wald statistic} = \frac{(\hat{\beta}_1^{(p)} - \hat{\beta}_1^{(q)})^2}{\hat{\sigma}^2_{\hat{\beta}_1^{(p)} - \hat{\beta}_1^{(q)}}} \qquad (2.6)$$

$\hat{\beta}_1^{(p)}$ is the parameter estimate from the *p*th quantile regression model and $\hat{\beta}_1^{(q)}$ is the parameter estimate from the *q*th quantile regression model. In this case, the Wald statistic follows a χ^2 distribution with one degree of freedom. If there are *p* covariates in the models, the Wald statistic follows a χ^2 distribution with *p* degrees of freedom (Koenker & Machado, 1999). Thus the Wald statistic can be readily extended for more complicated models for test of equivalence of coefficients across quantiles.

Obviously, Wald statistics can also be used to test the difference between a restricted model and an unrestricted model where one of them is nested within the other with only a subset of covariates. Likelihood ratio test can also provide similar information for this type of linear test. Koenker and Machado (1999) prove that these two tests are equivalent and follows a chi-square distribution under the null hypothesis.

Wald test based on the bootstrapping samples can be realized in at least two computer programs. Stata uses the sqreg command and Quantreg (Koenker, 2009), a free R package, uses the command anova.rq to test the equivalence of coefficients across quantiles. A goodness-of-fit test, Khmaladez Test in Quantreg (to be explained in the next section), provides additional criterion that can serve the same purpose. SAS has a command to test null hypothesis of $H_0 : \beta_1^{(p)} = 0$ but makes no mention of test of coefficients equivalence across quantiles. However, the same approach from Quantreg might be developed by manipulating commands already extant in SAS.

QRM Goodness-of-fit Index

In MRM, R^2 is the usual measure of goodness-of-fit. It is defined as:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \qquad (2.5)$$

This equation means R^2 is the ratio between the sum of squares due to regression and the sum of squares of the total model. In another word, it represents the proportion of variance in the response variable being explained by the covariates in the regression model. It ranges between 0 and 1 with a higher value indicating better fit.

In QRM, a similar index is suggested by Koenker and Machado (1999) which is the likelihood ratio between the sum of weighted absolute distances for the full *p*th quantile regression model and the sum of the weighted absolute distances for a model with only the intercept. Let $V^1(p)$ represents the former and $V^0(p)$ the latter, still modeling one covariate, the equation representation for this index is

$$R(p) = 1 - \frac{V^1(p)}{V^0(p)} = 1 - \frac{p \sum_{y_i \geq \hat{y}_i} \left| y_i - (\beta_0^p + \beta_1^p) \right| + (1-p) \sum_{y_i < \hat{y}_i} \left| y_i - (\beta_0^p + \beta_1^p) \right|}{p \sum_{y_i \geq \hat{Q}^{(p)}} \left| y_i - \beta_0^p \right| + (1-p) \sum_{y_i < \hat{Q}^{(p)}} \left| y_i - \beta_0^p \right|} \qquad (2.6)$$

Stata named this “pseudo- R^2” to distinguish it from R^2 for regular MRM.

For the model that only includes an intercept, the intercept is the sample *p*th quantile $\hat{Q}^{(p)}$ of the response variable. Both $V^0(p)$ and$V^1(p)$ are nonnegative since they are the sum of some absolute values. $V^1(p)$ is always equal or smaller than $V^0(p)$ since a covariate is supposed to have explained some variance. Thus R(*p*) is also within the range of [0,1], with a larger R(*p*) indicating a better model fit just like the R^2 .

The R(*p*) defined above naturally leads to a relative term that can be used to evaluate improvement in model fit by a more constrained model. Let $V^2(p)$ be the sum of the weighted absolute distances for the less constrained model and $V^1(p)$ for the more constrained one,

$$\text{Relative R}(p) = 1 - \frac{V^2(p)}{V^1(p)} \qquad (2.7)$$

However, this pseudo-R^2 index works well just with local comparisons at the same quantile. In reality, a covariate may have an effect on the response variable at the tails (p=.9 or .1) but not at the median (p=.5). Also, classical inference statistics depend on a distribution form, which destroys the advantage of QRM models that is distribution free for the error term (Koenker & Machado, 1999; Koenker & Xiao, 2002). For this purpose, a distribution free test, Khmaladez test, is developed. This test uses the martingale approach which was first developed by Khmaladez (1981) and was extended to QRM by Koenker and Xiao (2002). This test can test the covariate effect in location shift, location and scale shift or additional shape shift. It can test individual covariate effect as well as overall model effect, effect at a certain quantile or across all quantiles. This test function is available in the free R package Quantreg and the critical values were developed and summarized into a table by Koenker and Xiao (2002). Results of Khmaladez test also offers supplements to other measures of location, scale and skewness shifts such as the ones recommended by Hao and Naiman (2007).

QRM Coefficients as Measures of Location, Scale and Skewness Shifts

An important advantage of QRM over traditional MRM models is that QRM is robust to location, scale and skewness shifts as it does not require the normality and homoscedasitity assumptions like MRM does. On the other hand, all these shifts can be studied and compared across quantiles to see the different effects of the same covariate.

Location shift refers to the shift in the measure of central tendency. This is an expected part for most regression analyses and is not a unique feature limited to QRM. However, in addition to this, QRM reveals shape shifts of the response distribution that MRM fails to capture. Shape shifts include both a scale shift and a skewness shift with the former referring to the change in the dispersion of response variables and the latter the change in the skewness of the response variable distribution.

Usually, graphics are used to inspect these shifts. For example, Koenker and Hallock (2001) used a set of box-plots to show the location and scale shifts in the annual compensation of chief executive officers. They also demonstrated how the various quantile regression lines reveal the skewness and skewness shifts of the Food Expenditure at different Household Income value using Engels' data of 1857. Others graphed the coefficients from various quantile regression models and interpret shape shifts accordingly (Buchinsky,1994; Hao & Naiman, 2007; Prieto-Rodriguez et al., 2008).

Location shift is readily reflected in the coefficient estimates.

$$
\begin{aligned}
& SCALE(y \mid x+1) - SCALE(y \mid x) \\
& = [Q^{(1-p)}(y \mid x+1) - Q^{(p)}(y \mid x+1)] - [Q^{(1-p)}(y \mid x) - Q^{(p)}(y \mid x)] \\
& = [\hat{\beta}^{(1-p)}(x+1) - \hat{\beta}^{(p)}(x+1)] - [\hat{\beta}^{(1-p)}x - \hat{\beta}^{(p)}x] \\
& = \beta^{(1-p)} - \beta^{(p)} \ \mathit{for} \quad p < .5
\end{aligned}
$$

Shape shifts, however, are not as obvious. Hao and Naiman (2007) contributed to the quantile regression literature by developing measures of shape shifts using the QRM coefficient estimates. In their own words: “These measures provide direct answers to research questions about a covariate’s impact on the shape of the response distribution” (p.5). However, they also pointed out that these measures are suitable for a model with only main effects.

Compared with the standard deviation as the measure of spread for normal distribution, Hao and Naiman (2007) suggested the *pth quantile range/ quantile-based scale measure* (QSC) / *interquantile range* (IQR) for skewed distributions.[6] In quantile regression models, every *p* relates to two sample quantiles: $\hat{Q}^{(1-p)}$ (the [1-*p*]th quantile) and $\hat{Q}^{(p)}$ (the pth quantile). The *p*th quantile range can be defined as $IQR^{(p)} = \left|\hat{Q}^{(1-p)} - \hat{Q}^{(p)}\right|$.[7] This quantity describes the range of the middle (1-2*p*) proportion of the distribution. For example, for p=.1, the interquantile range is between the 10th and 90th percentiles and describes the middle 80% of the distribution. When *p*=.25, the interquantile range becomes the interquartile range.

Any reference group and a comparison group can also be compared on scale change. For any fixed quantile of choice, an interquantile range can be found as $IQR_R = U_R - L_R$ for the reference group and $IQR_C = U_C - L_C$ for the comparison group. The difference between these two IQRs is a measure of scale-shift effect which Hao and Naiman (2007) named "*difference-in-differences*" scale-shift effect or $SCS^{(p)}$.

Scale-shift effect or $SCS^{(p)}$ can be estimated by QRM coefficient estimates. For the same model used as the example in this chapter, $\hat{\beta}_1^{(p)}$ is the fitted coefficient for a covariate in a *p*th quantile-regression model and it indicates the change in any particular

[6] These various terminologies appear at difference places in Hao and Naiman (2007) but refer to the same thing.

[7] I made a change to the original equation by adding the absolute value constraint so that the sign will not interfere with interpretation. This is because if p<.05, IQR is positive, if p>.05, IQR is negative. Using absolute difference will always give the range in the positive term. Another way to define it is that $IQR^{(p)} = \hat{Q}^{(1-p)} - \hat{Q}^{(p)}$ when p < .5 and $IQR^{(p)} = \hat{Q}^{(1-p)} - \hat{Q}^{(p)}$ when p > .5 which is used for the derivation for SCS later.

quantile as the covariate increases by one unit. As a result, the corresponding *p*th interquantile range changes by the amount $\hat{\beta}_1^{(1-p)} - \hat{\beta}_1^{p}$, which is $SCS^{(p)}$.[8] The connection is easy to see:

$$\begin{aligned} SCS^{(p)} &= IQR_C^{(p)} - IQR_R^{(p)} = (Q_C^{(1-p)} - Q_C^{(p)}) - (Q_R^{(1-p)} - Q_R^{(p)}) \\ &= (Q_C^{(1-p)} - Q_R^{(1-p)}) - (Q_C^{(p)} - Q_R^{(p)}) \\ &= \hat{\beta}_1^{(1-p)} - \hat{\beta}_1^{p} \quad for \quad p < .5 \end{aligned}$$

$$\begin{aligned} SCS^{(p)} &= IQR_C^{(p)} - IQR_R^{(p)} = (Q_C^{(p)} - Q_C^{(1-p)}) - (Q_R^{(p)} - Q_R^{(1-p)}) \\ &= (Q_C^{(p)} - Q_R^{(p)}) - (Q_C^{(1-p)} - Q_R^{(1-p)}) \\ &= \hat{\beta}_1^{p} - \hat{\beta}_1^{(1-p)} \quad for \quad p \geq .5 \end{aligned}$$

Zero SCS gives evidence of no scale change. Positive value indicates increase in scale as the covariate value increase. Negative value indicates decrease in scale as the covariate increase.

A quantile-based skewness (QSK) measure is also proposed by Hao and Naiman (2007) compared to the traditional skewness measure for a normal distribution. This is the ratio of the upper spread to the lower spread relative to the median minus 1 or

$$QSK^{(p)} = \frac{Q^{(1-p)} - Q^{(.5)}}{Q^{(.5)} - Q^{(1-p)}} - 1 \qquad \text{for p<.5.}$$

Positive QSK indicates a positively skewed distribution and negative QSK indicates a negatively skewed distribution.

[8] Hao and Naiman (2007) pointed out, the scale effect does not depend on the reference group if only a linear QRM with no covariate interactions is fitted. When interactions exist, measure of SCS is more complex. As of now, interaction terms have not been studied extensively in quantile regression literature.

Along the same logic, difference in skewness between a comparison and a reference group is proposed as the "*ratio of the ratios*" or

$$\frac{\dfrac{(Q_C^{(1-p)} - Q_C^{(.5)})}{(Q_R^{(1-p)} - Q_R^{(.5)})}}{\dfrac{(Q_C^{(.5)} - Q_C^{(p)})}{(Q_R^{(.5)} - Q_R^{(p)})}}$$

where

$Q_C^{(1-p)} - Q_C^{(.5)}$ is the upper spread (range from the median to the upper quantile of interest) for the comparison group

$Q_R^{(1-p)} - Q_R^{(.5)}$ is the upper spread (range from the median to the upper quantile of interest) for the reference group

$Q_C^{(.5)} - Q_C^{(p)}$ is the lower spread (range from the median to the lower quantile of interest) for the comparsion group

$Q_R^{(.5)} - Q_R^{(p)}$ is the lower spread (range from the median to the lower quantile of interest) for the reference group

A value of 1 indicates no skewness shift. A value larger than 1 means the right-skewness is increased. A value less than 1 means the right-skewness is reduced. This value minus 1 gives the percentage change which is now called *skewness shift*, or SKS.

Hao and Naiman (2007) derived SKS from the QRM coefficient estimates in the following way. For this purpose, the typical covariate setting with only the intercept is defined as the reference group. The SKS for the middle 100(1-2*p*)% of the population is then:

$$SKS^{(p)} = \frac{\dfrac{Q_C^{(1-p)} - Q_C^{(.5)}}{Q_R^{(1-p)} - Q_R^{(.5)}}}{\dfrac{Q_C^{(.5)} - Q_C^{(p)}}{Q_R^{(.5)} - Q_R^{(p)}}} - 1 = \frac{\dfrac{\hat{\beta}^{(1-p)} + \hat{\beta}_0^{(1-p)} - \hat{\beta}^{(.5)} - \hat{\beta}_0^{(.5)}}{\hat{\beta}_0^{(1-p)} - \hat{\beta}_0^{(.5)}}}{\dfrac{\hat{\beta}^{(.5)} + \hat{\beta}_0^{(.5)} - \hat{\beta}^{(p)} - \hat{\beta}_0^{(p)}}{\hat{\beta}_0^{(.5)} - \hat{\beta}_0^{(p)}}} - 1 \qquad for \quad p < .5$$

SKS is a measure of skewness “above and beyond proportional scale shifts” (Hao & Naiman, 2007). A value of zero can mean either no scale shift at all or a proportional scale shift. However, a negative value means decrease in right-skewness due to the covariates and a positive value indicates increase in the right-skewness. Thus a non-zero value of SKS means a sure skewness shift.

One final note from Hao and Naiman is on the overall evaluation of a covariate’s impact on the inequality of the response. To decide whether all the shifts are statistically significant, it is necessary to examine the alignment of the signs of location, scale, and skewness shifts. For this purpose, two more terminologies are introduced: *in-sync* and *out-of-sync*. In-sync means that the signs for location, scale and skewness shifts are consistent with each other. Using comparison groups versus the reference group, if the signs are all positive, this means that the median of the comparison group is higher than the reference group, the scale of the comparison group is more spread about than the reference group and the comparison group is more right-skewed than the reference group. In the context of an inequality study, these in-sync shifts “make the total distribution more unequal” which means the covariate/ predictor “exacerbates inequality through both location and shape changes” (Hao & Naiman, 2007). On the other hand, when the shifts are out-of-sync, the covariate affects the location and shape of the response distribution in different directions which compromises the total effect due to this covariate.

Hao & Naiman (2007) provided examples of codes in STATA to calculate the location and shape shifts. Quantreg package in R produces Khmaladez test results. A combination of these two methods will reveal a complete picture of the stochastic relationship between variables that MRM cannot cover. However, both statistics are suitable for models without interaction terms. When interactions are involved, the approach by Hao & Naiman is not appropriate. It is unclear whether Khmaladez test statistics will be misleading when interactions terms are included in the model.

QRM Applications in the Educational Field

For a long time, estimation challenges precluded QRM use with large scale applications. Linear programming and modern technology makes efficient computation a manageable task. QRM has now become a standard function in programs such as STATA and SAS. As the result, QRM has become a common tool in many fields such as economics, finance, medicine, biology and environmental studies. QR application in the educational field was very recent. Most of these were on equality issues and appeared in journals of economics (Haile & Nguyen, 2008; Levin, 2001; Prieto-Rodriguez, Barros & Vieira, 2008; Wößmann, 2005). For example, Haile and Nguyen (2008) studied the achievement gap between ethnic groups and genders. Results from mean regression were consistent with traditional findings that Asian students scored better than White students in mathematics. Quantile regression results showed, however, that Asian *male* students performed better than their White counterparts only at the 0.1 quantile but Asian *females* outperformed their White counterparts at all quantiles except 0.1. Wößmann (2005) explored the heterogeneity in the central examination effect using several international databases and concluded that central exit examinations were conducive to students' performance and could even reduce the effect of parents' educational background. Levin's (2001) research on class size was also motivated out of economic concerns because smaller class size requires more resources. Using QR, he found out that despite conventional belief, reduced class size actually related negatively with scholastic achievement for less able students. Rather than the class size, peer effect was the variable in his data that correlated positively with students' achievement at the lower distribution of abilities. All these applications of QR, however, were closer to the economic field than the educational field. The single application that can be regarded as for the educational measurement and assessment purpose was Betebenner's (2009) growth model, named Student Growth Percentiles (SGP).

SGP builds on a simple concept to take prior status into consideration when modeling the current growth. This is necessary because it is not reasonable to expect all the students to grow at the same rate and achieve the same proficiency criterion within the same time frame (Betebenner, 2009). For example, a student who was at the 10th percentile in math assessment at grade one has a low probability to score at the 50th percentile at grade 2. However, another student who was at the 90th percentile at grade one is almost ensured to score above the 50th percentile the next year. Clearly, a reasonable description of growth should be based on conditional achievement. In Betebenner's model, growth is conditioned on prior status.

Traditionally, student's observed achievement scores were used to model the conditional growth (Chang et al., 2009; Galindo 2009; Han, 2008). However, the scale difference in scores across years makes the absolute mean difference between years not very informative (Betebenner, 2009; Yen 2007). As Yen (2007) and Betebenner (2009) showed, parents, teachers and administrators were not satisfied with how much students scored below, above or at a specific proficiency cut point. They wanted to know whether the score difference was appropriate and how far away the students were from being proficient compared to their peers. SGP takes advantage of quantile regressions and is aimed to show this relative or norm-reference achievement. As Betebenner (2009) points out, it serves as a supplement to mean-regression-based, criterion-referenced achievement.

SGP is a model that triggers the current research on the utility of quantile regression to study the language impact on math. Concept such as taking previous status into consideration is worthy of exploration. However, there are several issues with SGP. First, Betebenner' model does not consider the language impact in assessments. When language ability is confounded with content ability as literature has confirmed, partialling out the influence of language ability should be the first step before content achievement is modeled. Second, SGP is a full autoregression model in which all the previous scores were included in predicting the current score. For example, to model the 4th grade score, all first, second and third grade scores are counted in the model. This might magnify the influence of previous status too much. It is more reasonable to assume that wherever students start, their previous year's performance sufficiently serves as the base for current growth. The inclusion of all previous terms is more out of need for statistical fit rather than theoretical support[9]. To focus on students' achievement no matter what leads to it, attention should be directed to the current factors that test scores can reveal. In the current research, that factor is language.

Summary

Literature has confirmed the impact of language on math achievement. This impact is found for all population groups at various grades, for both genders, different ethnicity groups and groups with various social and family background. There was also evidence that the impact of language is not constant within grades or across grades. In addition, conclusions on the long-term effect were conflicting between studies. Limitations in research design and statistical techniques were the causes. To better explore the shifting influence of language on math achievement, longitudinal data together with quantile regression modeling are recommended. Advantages of quantile regression make it a more appropriate tool for these research questions. Available software facilitates the estimation and hypothesis testing. The introduction of quantile regression into the educational field has already started and this research aims for its application in the language research and assessment community. More specifically, quantile regression is the right tool to explore the differential relationship between language and math achievement.

[9] The author found out through personal communication with Betebenner that the decision to include scores from all previous years was mainly due to statistical needs. The model fitted best in this way.

Chapter III

DATA AND METHODOLOGY

Research Questions as a Review

Data and methodology were selected to answer the following four research questions:

1. How does language proficiency affect math achievement within and across grades?
2. How does math performance vary with respect to other background variables such as gender and socioeconomic status after language proficiency is controlled?
3. Does the math achievement gap between ELLs, former ELLs, and Non-ELLs increase or decrease as students move to higher grades?

Data

This research uses data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K). The ECLS-K was developed under the sponsorship of the U.S. Department of Education, Institute of Education Sciences and National Center for Education Statistics (NCES). Westat and Educational Testing Service

(ETS) conducted the actual study. The ECLS-K is an unprecedented study in terms of following a cohort of students from kindergarten to eighth grade to measure the cognitive and social development of children. It involves a seven-wave, multi-stage data collection design starting from Fall 1998. The baseline population includes 21,260 nationally representative kindergarteners at the beginning of the study from 934 public schools and 346 private schools throughout the country. However, minority, low-income, disabled, and special-needs children were oversampled. The sample was freshened in Spring 2000 to include children who were not in kindergarten neither full-time nor part-time in the US during the 1998-1999 school year. Thus, the ECLS-K data is representative of kindergartners in 1998 and first graders in 2000 in US school system but not for other grades. In addition, 30 percent subsamples were drawn in Fall 1999 for the purpose to measure the extent of summer learning loss. Table 3.1 is a replication from a user's manual to show the seven rounds of data collection time points, grade and school year.

Table 3.1 Crosswalk between Round Number of Data Collection, Grade, and School Year: School Years 1998-99, 1999-2000, 2001-02, 2003-04, 2006-07

Round number	Child Code	Grade	School year
1	C1	Fall-kindergarten	Fall 1998
2	C2	Spring-kindergarten	Spring 1999
3	C3	Fall-first grade (subsample)	Fall 1999
4	C4	Spring-first grade	Spring 2000
5	C5	Spring-third grade	Spring 2002
6	C6	Spring-fifth grade	Spring 2004
7	C7	Spring-eighth grade	Spring 2007

The complete dataset from K to 8th grade is described but only the last four time points are used for the core analysis in this research. There are three reasons for this decision. First, the first two time points are for kindergarten students. Kindergarten constructs can be reasonably assumed to be different from later years of schooling. The 8th grade psychometric report pointed out that some basic reading skills for kindergarten were not tested later because almost all children had mastered them by the end of the first grade. Second, the sampling framework changed at the fourth time point. Third, the last four time points correspond to grades 1, 3, 5 and 8. They are commonly of interest for assessment. These students were all tested in spring thus summer learning loss (if there is any) would not complicate the interpretation.

There are several data sets available. The specific set used in this study is the Kindergarten-Eighth Grade Full Sample Public-Use Data File especially prepared for longitudinal studies. This file includes data from the base, first-, third-, fifth, and eighth-grade years and have all data for all ECLS-K sample cases that have been publicly released in any of the rounds. More information about this dataset can be located in the Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and the Electronic Codebooks (Tourangeau, et al., 2009, hereafter referred to as the "Manual") and the ECLS-K Psychometric Report for the Eighth Grade (NCES 2009-002) (Najara, Pollack, & Sorongon, 2009).

Chapters five and ten in the Manual provide a different set of complete records other than the actual data file on the website. These are listed in Table 3.2. All the numbers reported in the manual are greater than the real data because the "complete" data in the manual include children who were excluded from direct assessment due to a disability. The real assessment data show these children as missing thus giving a smaller number for the record counts. Round 7 has much lower response rate because a new process was involved by asking parents' consent before collecting data. Also, children were given the choice to decide to participate or not. So, the big decrease in Round 7 was unavoidable. Round 3 is the subsample described above. Because there are several missing data codes that are not useful for this study, these are recoded by ignoring the reason for missing and replaced with a "." for this research. This procedure further decreases the count of actual observations. The corrected empirical valid counts are included in Table 3.2. All analyses are based on the final recoded data.

Table 3.2 Comparison of Direct Child Assessment Records: School Years 1998-99, 1999-2000, 2001-02, 2003-04, 2006-07

		Count		
Round	Grade	Manual	Data	Recoded
C1	Fall-kindergarten	19172	19126	17622
C2	Spring-kindergarten	19967	19917	18937
C3	Fall-first grade (subsample)	5291	5267	5053
C4	Spring-first grade	16727	16683	16336
C5	Spring-third grade	14470	14415	14280
C6	Spring-fifth grade	11346	11294	11265
C7	Spring-eighth grade	9358	9307	9225

There are various instruments used to collect data including direct child assessment, and questionnaires from children, parents and teachers. For the current study, direct reading and math assessments are used to indicate children's academic achievement and questionnaires are used to locate background variables such as gender, ELL status, socioeconomic status and race-ethnicity.

Reading scores and math scores are time-varying variables. Both of these are scale scores based on three-parameter IRT models. Reading scores are the independent variable and math direct assessment scores are the dependent variable. Psychometric quality is assumed good with ECLS-K data because of all the resources and experts involved. Also, the psychometric reports document the quality control procedures in great detail. The reading and math ability score reliability are all high with the lowest being .87. These values are replicated in Table 3.3.

Table 3.3 Reliability of Direct Assessments

Reliability	Round 1 K	Round2 K	Round 3 Grade 1 subsample	Round 4 Grade 1	Round 5 Grade 3	Round 6 Grade 5	Round 7 Grade 8
Reading	.92	.95	.96	.96	.94	.93	.87
Math	.91	.93	.94	.94	.95	.95	.92

Ideally, academic language proficiency should be used as the independent variable. As distinguished before, that is the part of language proficiency that continuously affects academic learning. However, this information is not available in ECLS-K. Reading ability is a proxy for language proficiency because understanding written text is the first form of language proficiency relevant to cognitive functions (Mestre, 1988).

Time invarying variables include ELL status, gender (GENDER), socioeconomic status (SES) and race-ethnicity. These variables are chosen based on past literature. GENDER is recoded ignoring reason for missing. ELL status is recoded to subset students into three groups: Non-ELLs (0), ELLs until Round 4 or Former ELLs (1) and ELLs after Round 4 (2)[10]. Because ELL status is based on an oral test, it reflects the English oral proficiency of students. This is different from the reading scores which reflect the reading proficiency of students. Thus in this research, two out of the four domains of language skills are actually represented.

SES is a composite of parents' income, educational levels and occupations. It involves 13 income categories, nine educational levels and 22 occupation types and is a weighted z-score. In other studies, there were usually either a high- or a low-SES group (Abedi et al 2005; Tate, 1997) or an above- or below-Bachelor's degree educational group (Abedi et al 2005; DOE, 2008). Because SES is a continuous variable in this study, it should act as a more powerful covariate than in the previous research.

Race is the variable that classifies students into different ethnic groups. It is based on parent reported data. It includes traditional categories of White, African American, Hispanic and Asian. Some other groups are recoded into two more categories: Isolated (Native Hawaiian or other Pacific Islander, American Indian or Alaska Native) and Others (more than one race specified). Because Others does not represent a meaningful comparison group for my purpose, it is not used for any analysis.

There is balanced proportion of the two genders (10950 boys and 10446 girls), but a strikingly small portion of ELLs and Former ELLs (2576 or 14.9%). This might be explained by the fact that many ELLs are immigrants who are not born in the USA and

[10] Language screening tool used to decide student ELL status was not used after round 4 because most students are regarded as having met the oral proficiency by then.

move here much later than kindergarten or first grade. The cohort of this study, on the other hand, was decided at kindergarten and/or first grade. As the manual explained, the sample does not represent the current student population at each grade level in the U.S.

Descriptive statistics reveal that the unconditional math and reading scores both changed in location and shape across time. Appendix A shows the distribution of reading and math score at the last four rounds of data collection. For every grade, there is a box-plot, a quantile plot and a histogram. Figure 3.1 shows the residual distribution from math scores by mean regression with all the covariates controlled. The graphs clearly show violation of normality and homoscedasticity assumptions for mean regression modeling. This justifies the choice of quantile regression over mean regression as the statistical technique for this data.

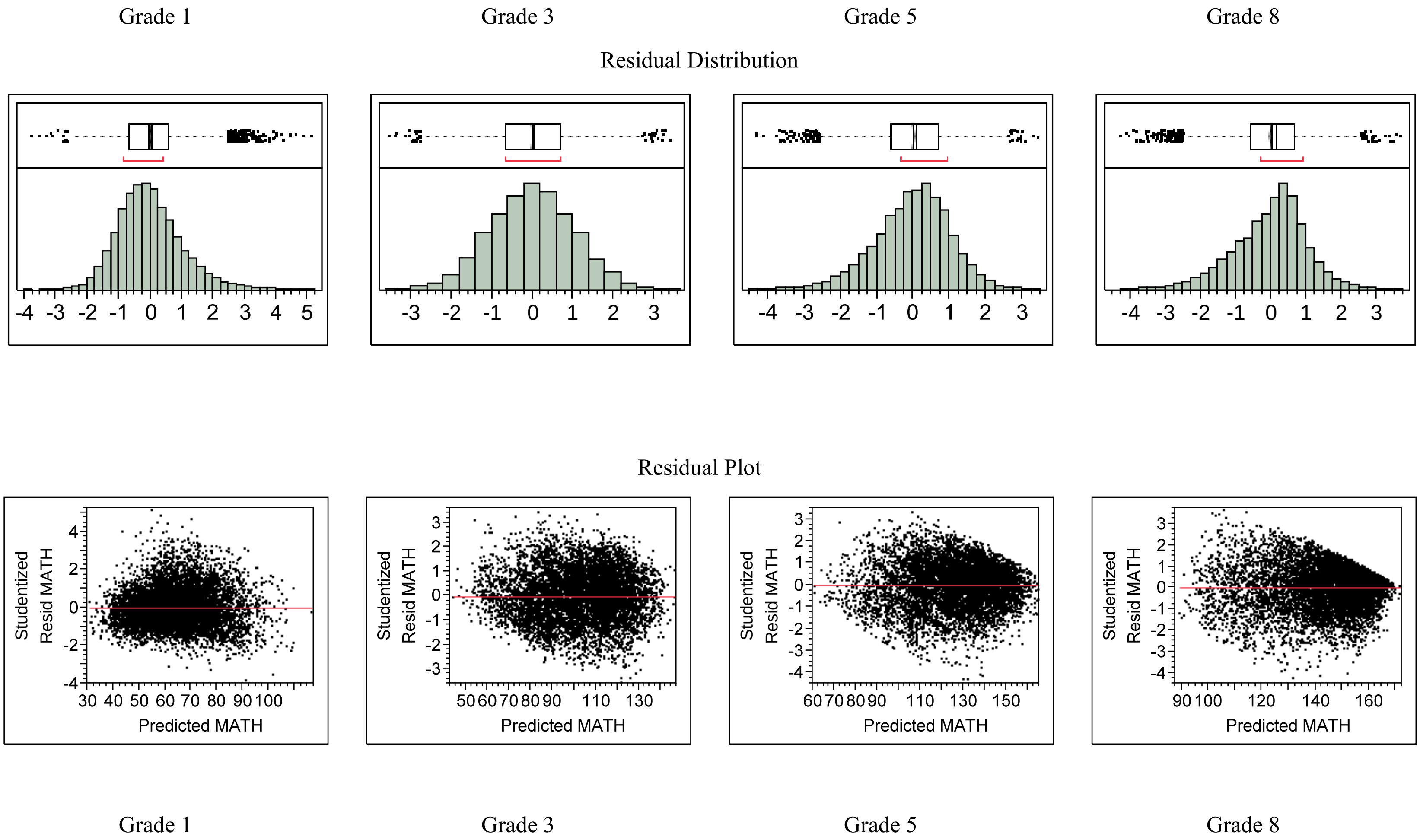


Figure 3.1 Residual Distribution and Residual Plot against Fitted Math

Models

The relationship between language and math is modeled through regressions. A linear mean regression serves as the base model. Seven quantile regressions are studied with detail per grade to detect the differential language effect[11]. The seven quantiles are 5th, 10th, 25th, 50th, 75th, 90th and 95th. Finer differences in quantiles are used at both ends because it is suspected that the impact may vary more at ends than in the middle. For example, if Ausubel and Robinson (1969) were right, language may have smaller impact on math achievement at lower ability level than at the higher ability level.

The mean regression is represented as

$$y_i = \beta_0 + \mathbf{X}_{im}\boldsymbol{\beta}_m + \varepsilon_i \qquad (3.1)$$

i refers to the person and *m* refers to the specific independent variables. **X** and **β** are vectors representing the set of independent variables and corresponding coefficients. Independent variables in this study include ELL status, gender, SES, race-ethnicity and reading (proxy for language proficiency). Reading scores are group mean centered at each grade to facilitate interpretation of the intercept. Appendix A summarizes the names and scale of measurement of all the variables. To facilitate understanding, the full equation is written out below in the form of a mean regression.

[11] For graphic presentations, more quantiles are used. The APPENDIX C graphics are based on 50 quantile points and figures in Chapter 4 are based on 19 quantile points evenly spread between .05 to .95.

$$\begin{aligned}\mathrm{MATH} = {} & \mathrm{READING} + \mathrm{SES} + \mathrm{GENDER} + \mathrm{FORMER} + \mathrm{ELL} + \mathrm{BLACK} + \mathrm{HISPANIC} + \mathrm{ASIAN} + \mathrm{ISOLATED} + \\ & \mathrm{READING} * \mathrm{SES} + \mathrm{READING} * \mathrm{GENDER} + \mathrm{READING} * \mathrm{FORMER} + \mathrm{READING} * \mathrm{ELL} + \\ & \mathrm{READING} * \mathrm{BLACK} + \mathrm{READING} * \mathrm{HISPANIC} + \mathrm{READING} * \mathrm{ASIAN} + \mathrm{READING} * \mathrm{ISOLATED}\end{aligned}$$

It is necessary to distinguish between a variable and a regressor. A variable is of the original interest to the current study, a regressor is the actual manifestation that enters the regression equation. A variable can have several forms (regressors) such as the original value, a squared term, or an interaction. It can also be on any scale. To acknowledge these as well as the existence of measurement error, all the regressors are represented by capitalized letters. The original variables of interest are written in the traditional way which can be lower case (e.g. gender) or capitalized (e.g. SES). Reporting of regression results mostly involves the capitalized terms and interpretation involves the traditional form of the term.

Descriptive statistics show that there are very few students falling into several of the ELL-by-race-ethnicity categories. Interactions between ELL status and race-ethnicity group thus are not included in the current research although these are of interest. For parsimony reasons and to focus on the issue of reading and math achievement, interaction terms between SES and gender and SES and ELL status are also not included in the model. Three-way interactions are not explored because this will lead to four-way interaction in the longitudinal model where the variable TIME (grade) is added to the model. Interpretation of three-way interaction is already hard to translate into practical meaning (Good & Hardin, 2006). As a result, the highest interaction term explored throughout this study is three-way as in the longitudinal model.

Quantile regressions are represented by

$$y_i = \beta_0^{(p)} + \mathbf{X}_{im}\boldsymbol{\beta}_m^{(p)} + \varepsilon_i^{(p)} \qquad (3.2)$$

Again, *i* is the person, *m* is the specific independent variable. *p* refers to the specific quantile regression. For every grade, there are seven quantile regression models corresponding to the percentile value of .05, .1, .25, .5, .75, .90 and .95.

Both equations are replicated for grade 1, 3, 5 and 8. This means in total, there are 32 grade level regression equations to be solved. Also, since the categorical variables are all dummy coded to enable comparisons, every full equation produces 17 coefficient estimates. The large sample size makes estimation possible. Standard errors will be produced by bootstrapping with 1000 replications.

Besides separate models for each grade, an overall longitudinal model is proposed to include the time variable. The equation looks like 3.2 but with an extra “TIME” variable added to the right side. The reading scores are grand-mean centered across all grades. TIME is coded as 0,1,2 and 3 with 0 being the first grade. The differential language impact across time is revealed through the interaction term between TIME and READING. Similarly, two way interaction terms between TIME and gender, and TIME and SES will also reveal the shifting influence of gender and SES on math achievement across time. The same seven quantiles are decided to see how all these effects vary depending on the students’ conditional math ability in a continuum spread from 5th percentile to the 95th percentile and across all four grades.

Parameter estimates will be obtained using the Quantreg packages in R, Stata and SAS to cross check. All programs produce standard error estimates and relevant information to test significance but SAS does not have a ready function to test equivalence of coefficients between quantiles. Wald statistics will be used to make the final conclusion of the significance of the coefficients. Wald test, likelihood ratio test and the pseudo-R^2 statistics as proposed by Koenker and Machado (1999) will be used as indices of model fit. Khmaladez test results (Koenker & Xiao, 2002) will detect the individual as well as the overall location shift or location-and-scale shift effect of the variables.

Whenever there are large amounts of information to present, graphics can help. Due to the many quantile models involved, graphics can supplement tables of statistics to facilitate reading. However, since graphics are supposed to highlight only key points, correct reading and full understanding of results will have to rely on tables as well. This is the dilemma of complex models: there will be more information, but there is more of a challenge to understand the information.

Chapter IV

RESULTS

The purpose of the current study is two-fold: to explore the changing relationship between language proficiency and math achievement and to demonstrate the advantages of quantile regression over regular mean regression. The purposes are well reflected in the following three research questions:

1. How does language proficiency affect math achievement within and across grades?
2. How does math performance vary with respect to other background variables such as gender and socioeconomic status after language proficiency is controlled?
3. Does the math achievement gap between ELLs, former ELLs, and Non-ELLs increase or decrease as students move to higher grades?

Based on literature and the availability of information in the data, the current research involves five independent variables, include reading scores (READING), gender (GENDER), SES, three ELL status (Non-ELL, FORMER and ELL) and five race-ethnicity groups (WHITE, BLACK, HISPANIC, ASIAN and ISOLATED). The longitudinal model also includes the grade (TIME). The dependent variable is math score

(MATH). Appendix A includes more descriptions of the variables and Appendix B provides relevant descriptive and diagnostic statistics.

In brief, the results show that language proficiency is the key variable that explains the math achievement gaps between ELLs and Non-ELLs. In addition, quantile regression revealed that language influence on math achievement differed between students with different math ability. The strength of relationship also decreases as students move up to higher grades. Consistent with past research, gender, SES, and race-ethnicity are all significantly related to math achievement. However, quantile regression revealed differential relationships depending on students' math ability.

Due to the large amount of and close relation among all results, the following sections are not organized to answer each research question separately. Rather, grade level models are discussed first and then longitudinal models second. Within each section, overall model fit is discussed before individual variables are interpreted. Grade level models serve as a base for longitudinal model building but results from the latter are interpreted with more confidence and detail since it has all the variables in the same model and was a statistically best-fitting model. Answers to research questions are summarized at the end based on key observations that stand out among all the results. A quick glance over the summary may guide readers to filter through the logic and arguments made here on.

Grade Level Results

Overall Model Fit

Three sets of regression models are compared and evaluated: a simple regression that includes only READING as the independent variable and two multiple regressions that include other covariate variables, with and without interaction terms. Table 4.1 summarized all the R^2s from mean regressions and the pseudo-R^2s from seven quantile regressions. R^2 revealed that READING by itself explains 44%-54% of the variance in MATH. Adding other covariates helps explain at most an additional 4% of variance. Adding interaction terms yields at most an additional 1% of the variance. Heuristically, Pseudo- R^2 from each quantile regression gives the same message as mean regressions. The absolute values are much smaller in Pseudo- R^2 than R^2 because the calculation is based on absolute deviation rather than squared deviation[12]. However, mean regression results show that about half of the variance in MATH is not explained by the dependent variables in all these models. The impact of this will be discussed later.

[12] $R^2 = 1 - \frac{SSE}{SST}$ and $R(p) = 1 - \frac{V^1(p)}{V^0(p)}$ .The ratio between the squared deviation of the hypothesis model and the null model, i.e. $\frac{SSE}{SST}$ is much smaller than the ratio between the absolute deviation of the two models ,i.e. $\frac{V^1(p)}{V^0(p)}$. Thus R^2 is larger than pseudo R^2.

Table 4.1
R-square for Mean Regression and Pseudo-R-square for Quantile Regressions

		Mean regression	Quantile regressions						
Grade	Model		.05	.1	.25	.5	.75	.9	.95
Grade 1	Simple regression								
	N=16334	.442	.299	.292	.272	.259	.265	.256	.225
	Multiple regression								
	N=8072	.468	.303	.292	.280	.278	.292	.287	.260
	With interactions								
	N=8072	.473	.307	.297	.286	.282	.292	.290	.267
Grade 3	Simple regression								
	N=14263	.543	.288	.317	.346	.342	.318	.291	.271
	Multiple regression								
	N=8023	.579	.318	.350	.376	.367	.346	.315	.297
	With interactions								
	N=8023	.580	.326	.355	.378	.367	.347	.319	.305
Grade 5	Simple regression								
	N=11256	.539	.342	.360	.362	.332	.288	.244	.203
	Multiple regression								
	N=7992	.580	.393	.403	.396	.358	.316	.269	.232
	With interactions								
	N=7992	.582	.395	.404	.396	.361	.324	.280	.246
Grade 8	Simple regression								
	N=9212	.535	.390	.385	.393	.332	.269	.195	.148
	Multiple regression								
	N=7959	.567	.417	.414	.393	.354	.288	.211	.163
	With interactions								
	N=7959	.568	.421	.416	.394	.356	.296	.224	.178

Note: Simple regressions include only Reading as the indicator. Multiple regressions include all the independent variables described in Chapter 2. The third model includes all independent variables and two-way interaction terms of interest.

Although model improvement beyond the main effect of READING in terms of R^2 is not big, Wald and Likelihood Ratio tests produce evidence that by including all other independent variables and interaction terms, both mean regression and quantile regression models fit the data better statistically (Appendix C). This is true for all grades.

The pattern of significance of each specific variable or interaction term is not the same across quantiles or between grades (Appendix D). To reach the statistically best-fitting models, insignificant terms should be dropped. However, this means the models at each grade will be different. For example, based on test statistics, READING*HISPANIC should be kept only for some lower quantile regressions at grade 1 but should be dropped completely from all quantile regressions at all other grades. ELL should be dropped from grade 1, 3 and 8 but should be kept at grade 5. To make things comparable, the same model is maintained for all grades, which includes the same main effects and interaction terms although not all of them are significant at all quantiles and/or all grades. Appendix D summarizes the significance test results for each variable and the interaction term. The details of the differences are discussed under the individual effects section below.

Quantile process plots (Appendix E) are used to present the slope estimates for the same variable at each quantile. These plots reveal that READING has a very strong and precisely estimated effect on MATH over the entire math score distribution. READING process plots do not mimic the shape of the intercept process plot, meaning READING exerts more than a location-scale effect on math score distribution. (Koenker

& Xiao, p.23). Also, several other slope coefficients do not stay constant which means a mean regression is not enough to describe these covariates' relationship with math scores. Table 4.2 summarizes the results of tests on the equivalence of the slope coefficients across quantiles. It gives additional statistical evidence (at $p \leq .05$) that the effects of READING and READING*SES varies across quantiles at all four grades. SES, GENDER, BLACK, READING*GENDER, and READING*BLACK have a differential effects at some of the grades. All other variables, that is, FORMER, ELL, HISPANIC, ASIAN, ISOLATED, READING*FORMER, READING*ELL, READING*HISPANIC, READING*ASIAN and READING*ISOLATED have a constant effect across quantiles at all grades. This means quantile regression does not give statistically more information than a mean regression on this last group of variables.

Khmaladez test based on main effects (Appendix F) shows similar results. There is statistically significant location and scale shift effect for all covariates and shape-shifting effect for some variables. Although not all individual covariates are related to the shape change in math score distribution by themselves, there is strong evidence that together, they do. They contribute to the shape shift in the math scores distribution for grade 1 (negatively skewed) and 8 (positively skewed).

Table 4.2.
Test of Equivalence

	Grade 1	Grade 3	Grade 5	Grade 8		Grade 1	Grade 3	Grade 5	Grade 8
	F(6, 8054)	F(6,8005)	F(6,7974)	F(6,7941)		F(6, 8054)	F(6,8005)	F(6,7974)	F(6,7941)
READING	7.73	8.67	59.94	104.41	READING*SES	3.49	6.86	4.83	7.09
	.00	.00	.00	.00		.00	.00	.00	.00
	***	***	***	***		***	***	***	***
SES	3.2	1.85	1.19	2.13	READING*GENDER	1.72	3.83	1.89	5.32
	.00	.09	.31	.05		.11	.00	.08	.00
	***	*		*			***	*	***
GENDER	13.26	6.44	.50	5.22	READING*FORMER	.65	1.24	.36	1.22
	.00	.00	.81	.00		.69	.28	.91	.29
	***	***		***					
FORMER	.50	.56	.24	1.47	READING*ELL	.14	.41	1.01	.83
	.81	.76	.96	.18		.99	.88	.42	.55
ELL	.42	.69	.72	.33	READING*BLACK	1.26	3.51	6.07	3.14
	.87	.66	.63	.92		.27	.00	.00	.00
							***	***	***
BLACK	6.14	1.88	2.25	.97	READING*HISPANIC	.97	.38	.49	1.45
	.00	.08	.04	.44		.45	.89	.82	.19
	***	*	**						
HISPANIC	1.35	.94	.27	.86	READING*ASIAN	.45	1.6	1.72	.9
	.23	.46	.95	.53		.84	.14	.11	.49
ASIAN	.64	1.18	1.06	.79	READING*ISOLATED	.55	1.64	1.77	1.37
	.70	.31	.38	.58		.77	.13	.10	.22
ISOLATED	1.05	.13	1.06	.46					
	.39	.99	.39	.84					

Note: First rows are the estimates, second row the p-value. *** indicates significance level at or below .01. ** indicates significance level at or below .05. * indicates significance level at or below .1. The .1 significance level is listed just for information. Decisions about significance are based at alpha level of .05 or lower.

Individual Variable Effects

To better examine the shifts of slope coefficients at each grade, individual coefficients are plotted with the same vertical scale range across grades. Figure 4.1 and 4.2 present the main effects and the interaction terms respectively. These figures are based on 19 quantile regressions for smoother appearance although all the relevant tables such as Appendix D include only seven quantiles for ease in reading.

For all the figures, x-axis represents the conditional quantiles under examination, ranging from .05 to .95. Y-axis corresponds to the intercept and individual slope estimates. The black solid line with shaded band is the coefficient estimate with confidence interval from quantile regression. The confidence band outside the quantile interval (.05, .95) is not displayed because there are usually insufficient data at extremes, making the confidence intervals not stable at those locations (SAS, 2008, p.5361). The red solid line is the mean regression estimate and the dashed lines are the upper and lower limit of the confidence interval. When possible, a vertical line at y=0 is also displayed. This corresponds to the null hypothesis of the parameter being zero.

The intercepts represent the conditional math score of a white, male, non-ELL student with average reading ability at the grade level and average socioeconomic status. The mean conditional math scores for the reference group are 66.4, 104.54, 129.05 and 145.81 for grade 1, 3, 5 and 8. The median (50th quantile) conditional math scores for the reference group are 64.9, 104.99, 130.65 and 147.42.

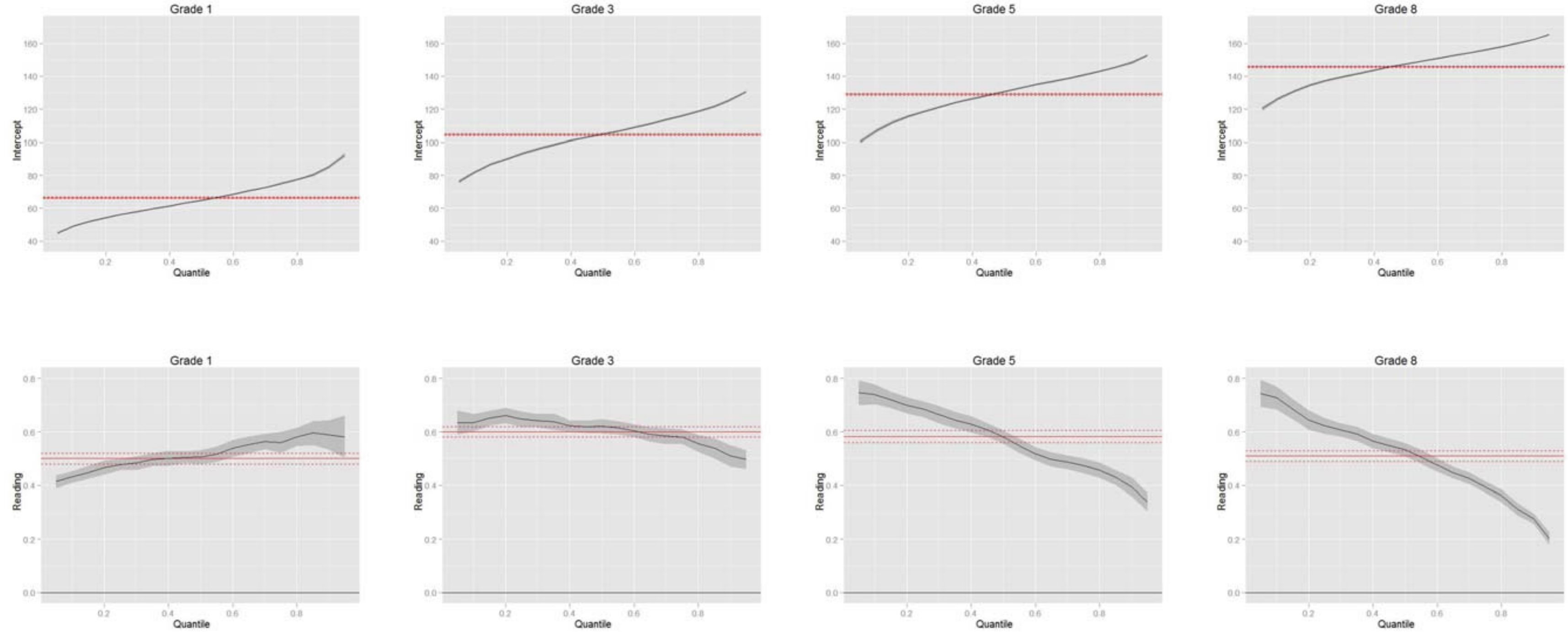


Figure 4.1a. Intercept and Main Effect READING

Note: Vertical scale range is (40,170) for intercept and (0, 0.8) for READING.

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.

The black line with shade area represents the quantile regression results with 95% confidence interval.

The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.

A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

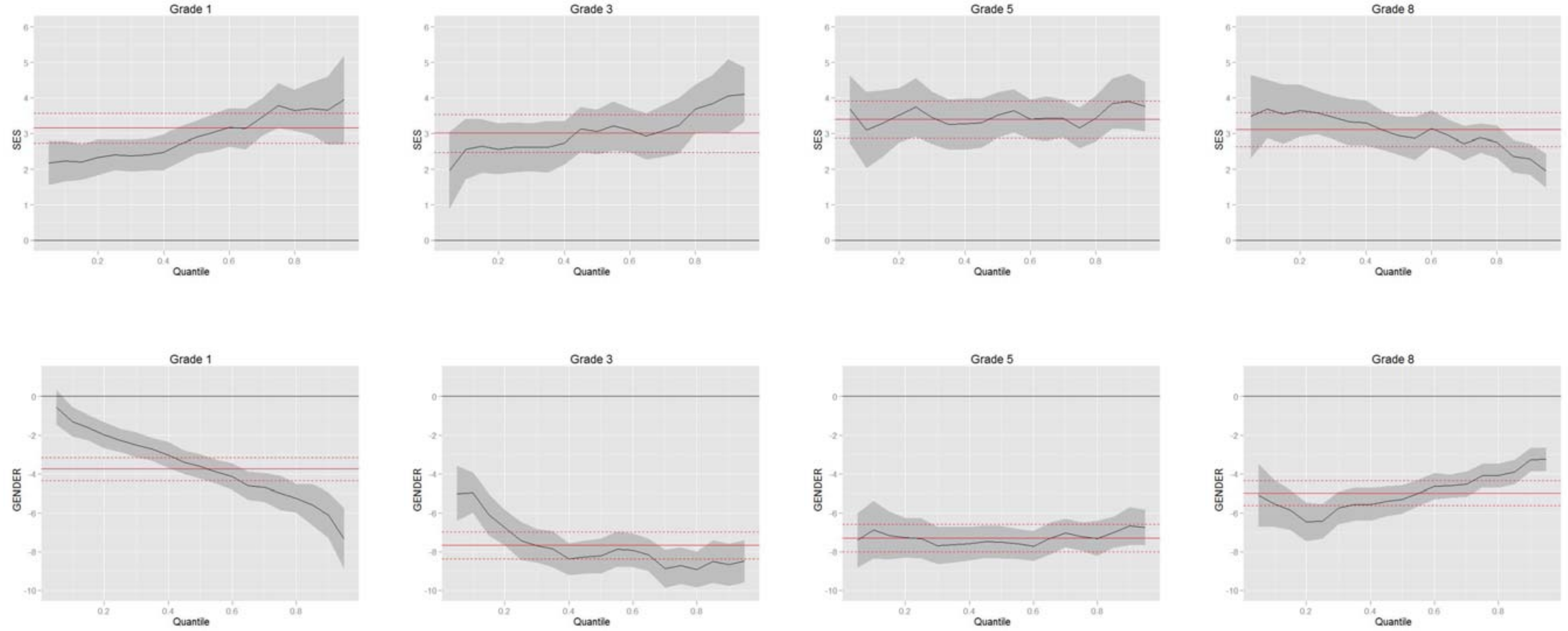


Figure 4.1b. Main Effect: SES and GENDER
Note: Vertical scale range is (0,6) for SES and (-10,1) for GENDER
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

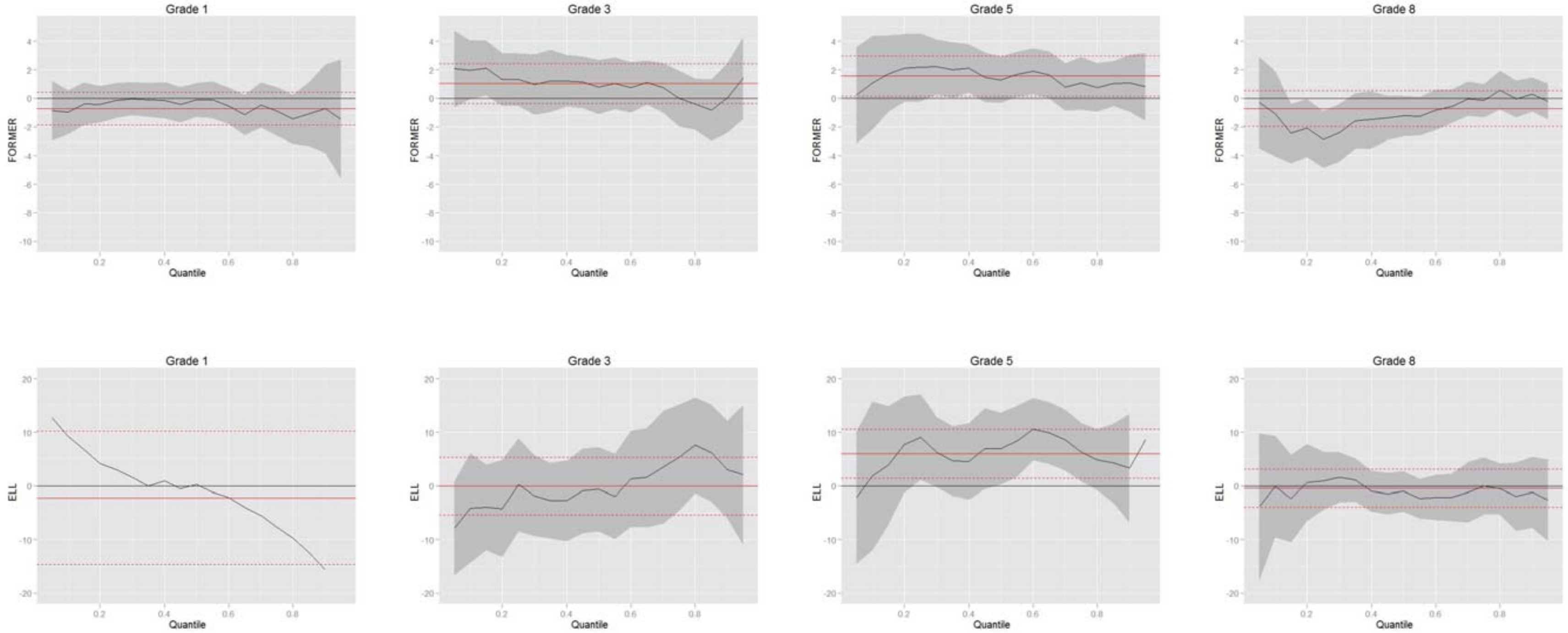


Figure 4.1c. Main Effect: Former ELL and ELL
Note: Vertical scale range is (-10, 5) for Former ELL and (-20,20) for ELL
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.
The shaded confidence intervals are not shown for ELL at Grade 1 and at some higher quantiles at Grade 5 because they are beyond the vertical range of (-10, 5).

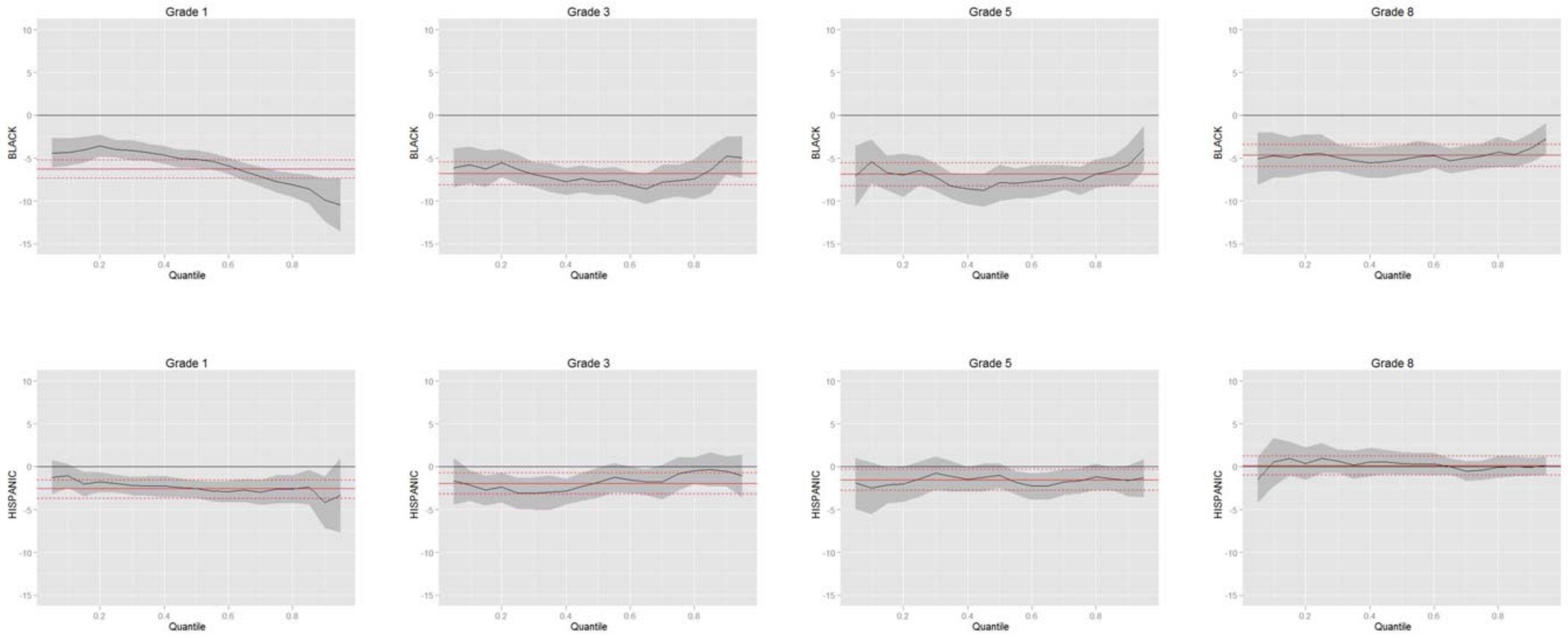


Figure 4.1d. Main Effect: BLACK and HISPANIC

Note: Vertical scale range is (-15, 10)

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.

The black line with shade area represents the quantile regression results with 95% confidence interval.

The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.

A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

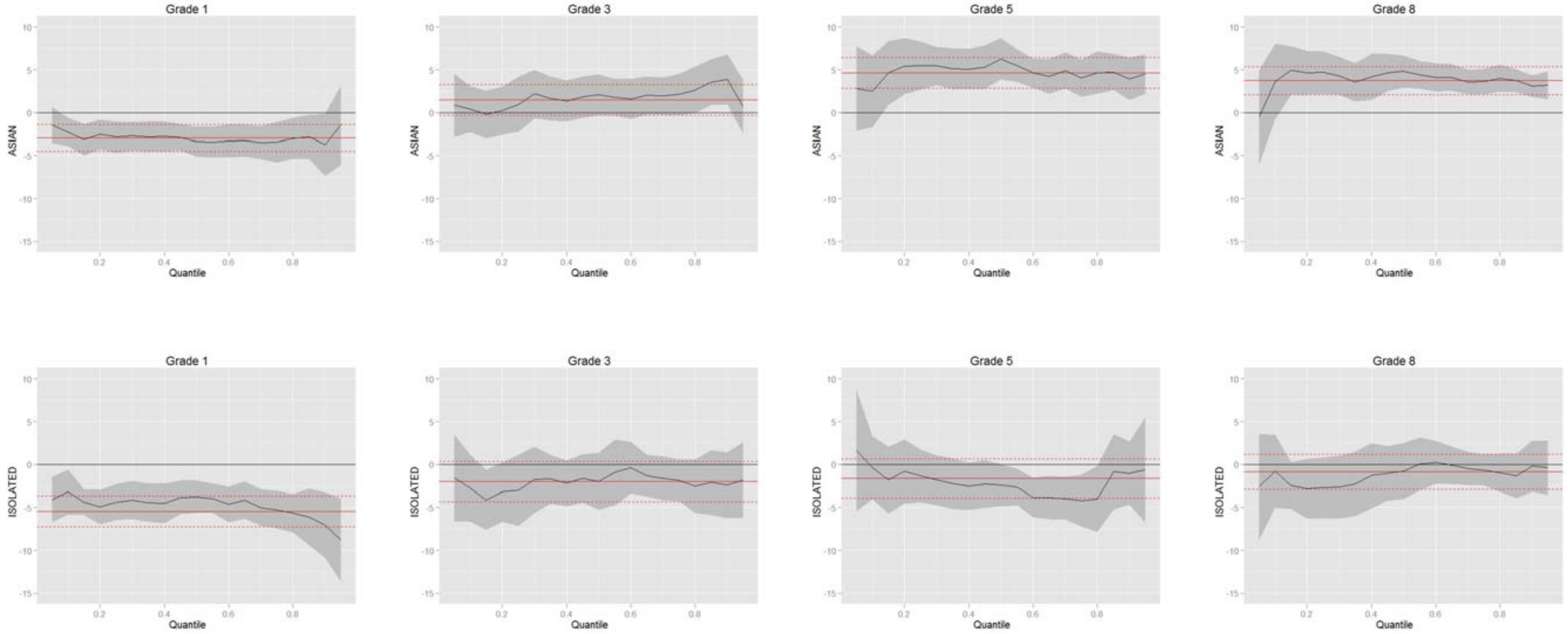


Figure 4.1e. Main Effect: ASIAN and ISOLATED
Note: Vertical scale range is (-15,10)
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

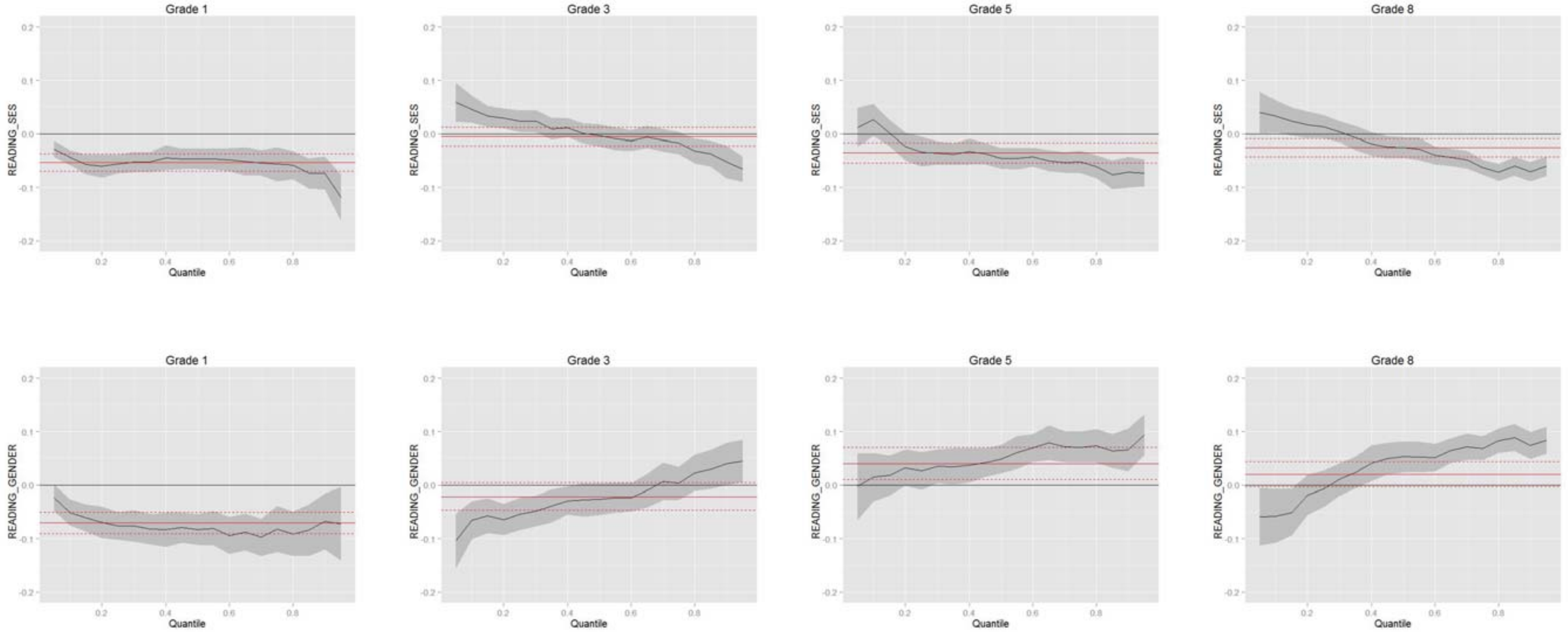


Figure 4.2a. Interaction: READING_SES and READING_Gender
Note: Vertical scale range is (-.2, .2)
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

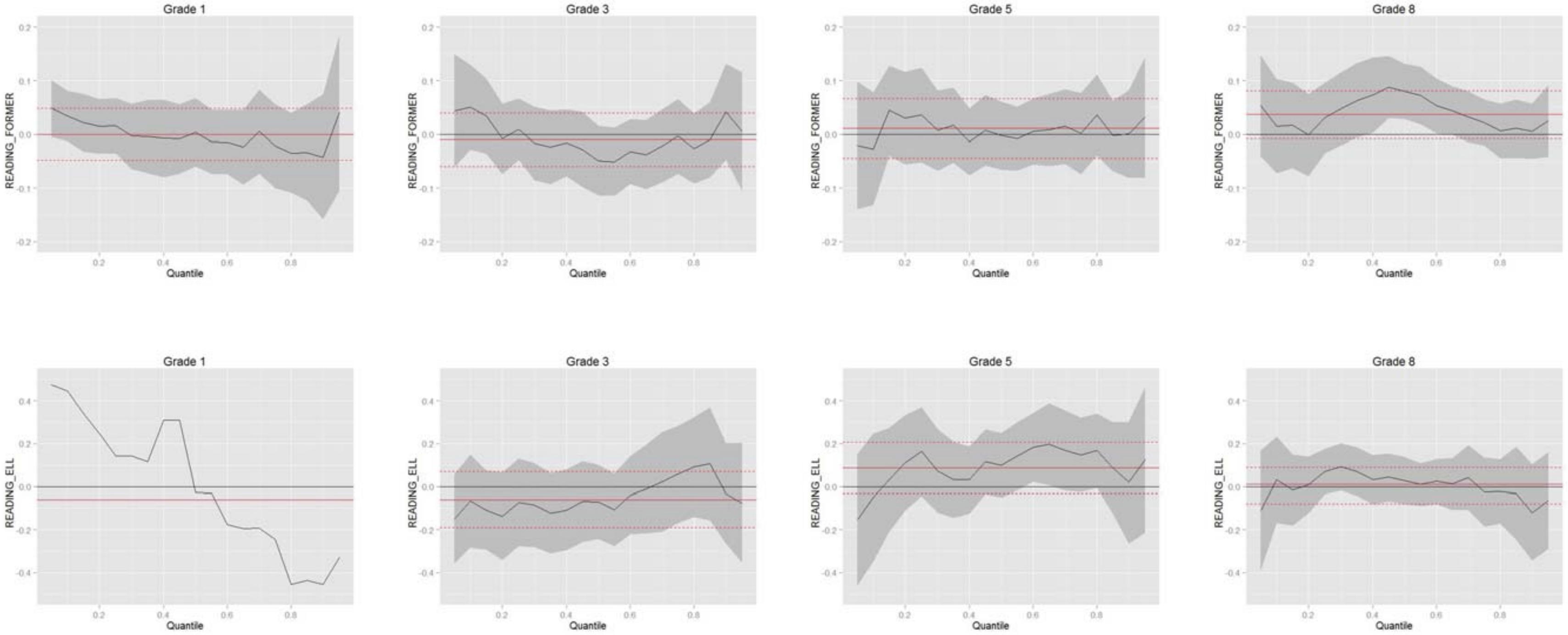


Figure 4.2b. Interaction: READING_FORMER and READING_ELL
Note: Vertical scale range is (-.2, .2) for READING_FORMER and (-.5, .5) for READING_ELL
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.
The confidence band is not shown for READING_ELL because it is out of the range of vertical scales used here.

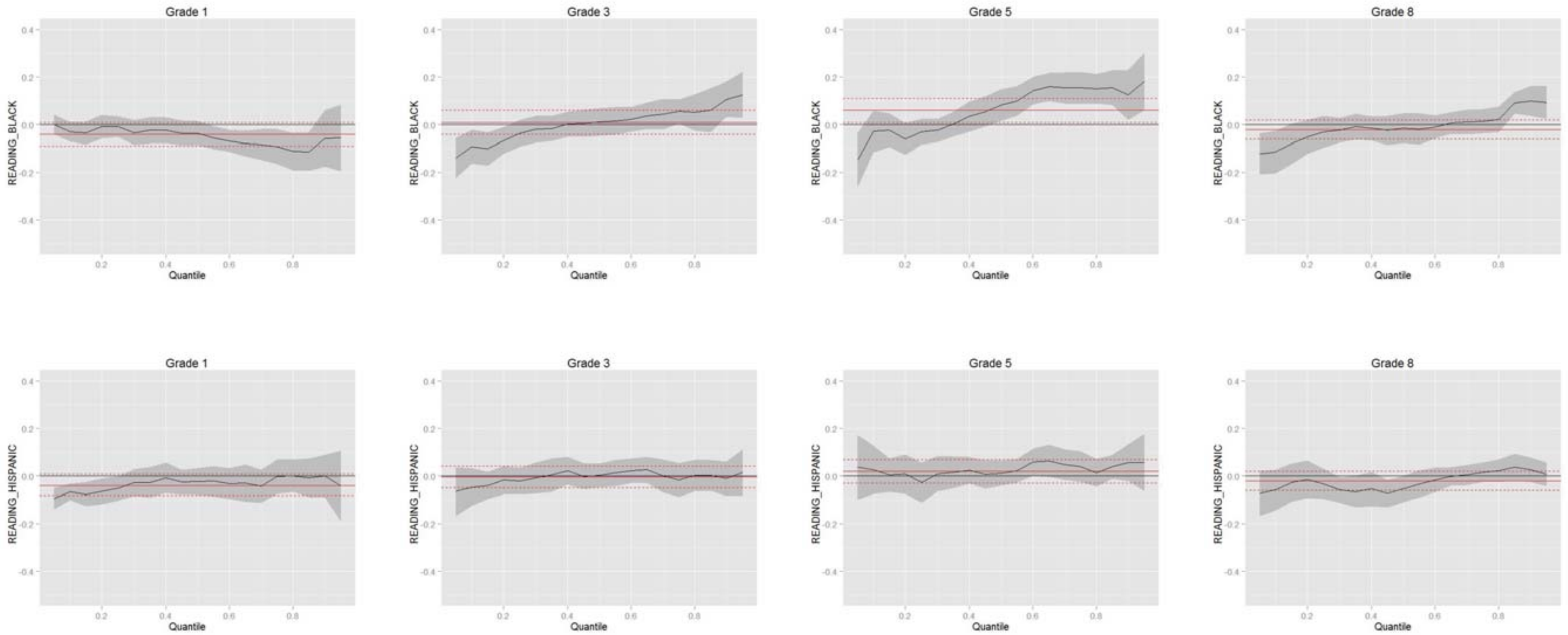


Figure 4.2c. Interaction: READING_BLACK and READING_HISPANIC
Note: Vertical scale range is (-.5, .4)
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

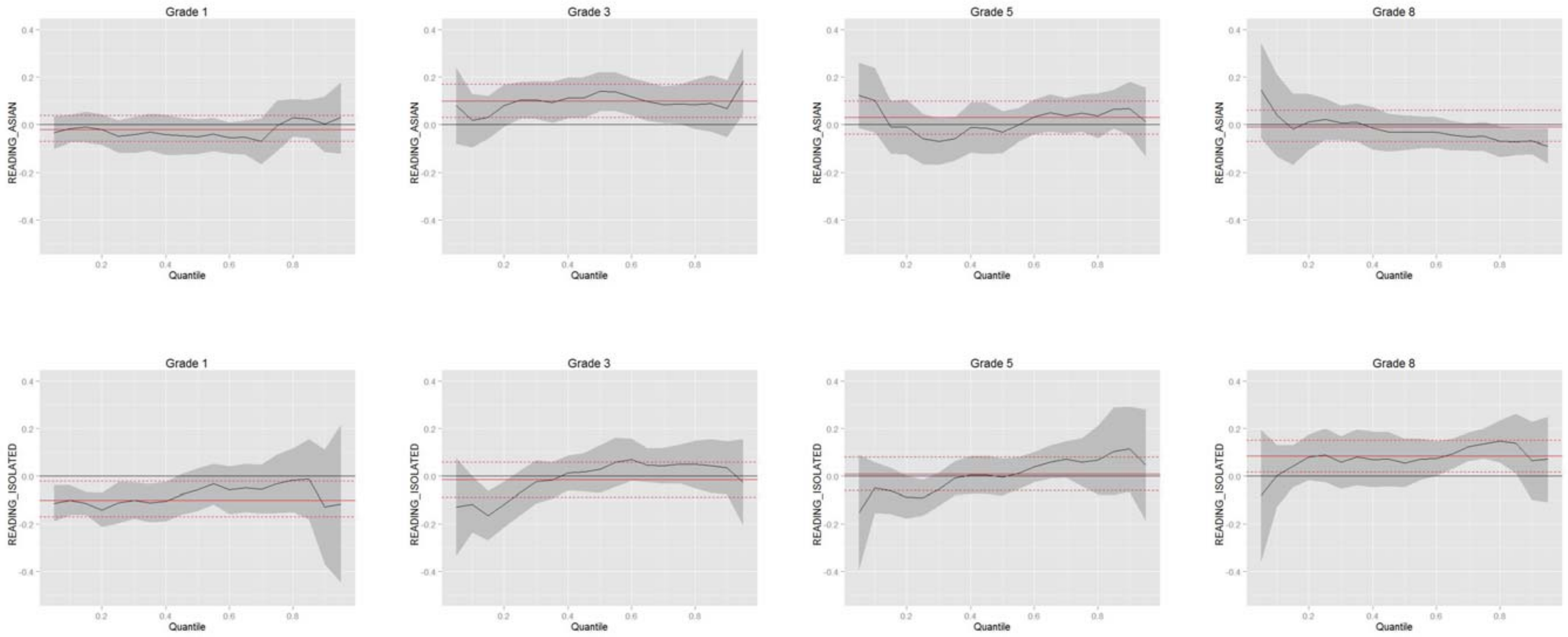


Figure 4.2d. Interaction: READING_ASIAN and READING_ISOLATED
Note: Vertical scale range is (-.5, .4)
The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

Several main effects immediately stand out. Both READING and SES are positively related to math scores across all quantiles at all grades. Both GENDER and BLACK are negatively related to math scores across all quintiles at all grades. More specifically, students with higher language proficiency have higher math scores; students with higher SES status have higher scores; girls have lower math scores than boys; and black students consistently scored lower than their white counterparts. There are significant interactions that moderate the magnitude of these main effects; however, they do not change the direction of these effects.

Although mean regression results detected the statistical significance of these four main effects on average, quantile regressions revealed that the same variable does not have the same effect along the conditional math ability distribution. At grade 1, as students' math ability increases, the relationship between language and math achievement also increases in strength (Figure 4.1a). Mean regression overestimates this relationship for low math ability students and underestimates it for high math ability students. This trend reversed from grade 3 to grade 8[13]. At these later grades, although language is still strongly and positively related to math scores, the strength of the relationship decreases as students' language proficiency increases. At these grades, mean regression underestimated the relationship for low math ability students and overestimated it for high math ability students.

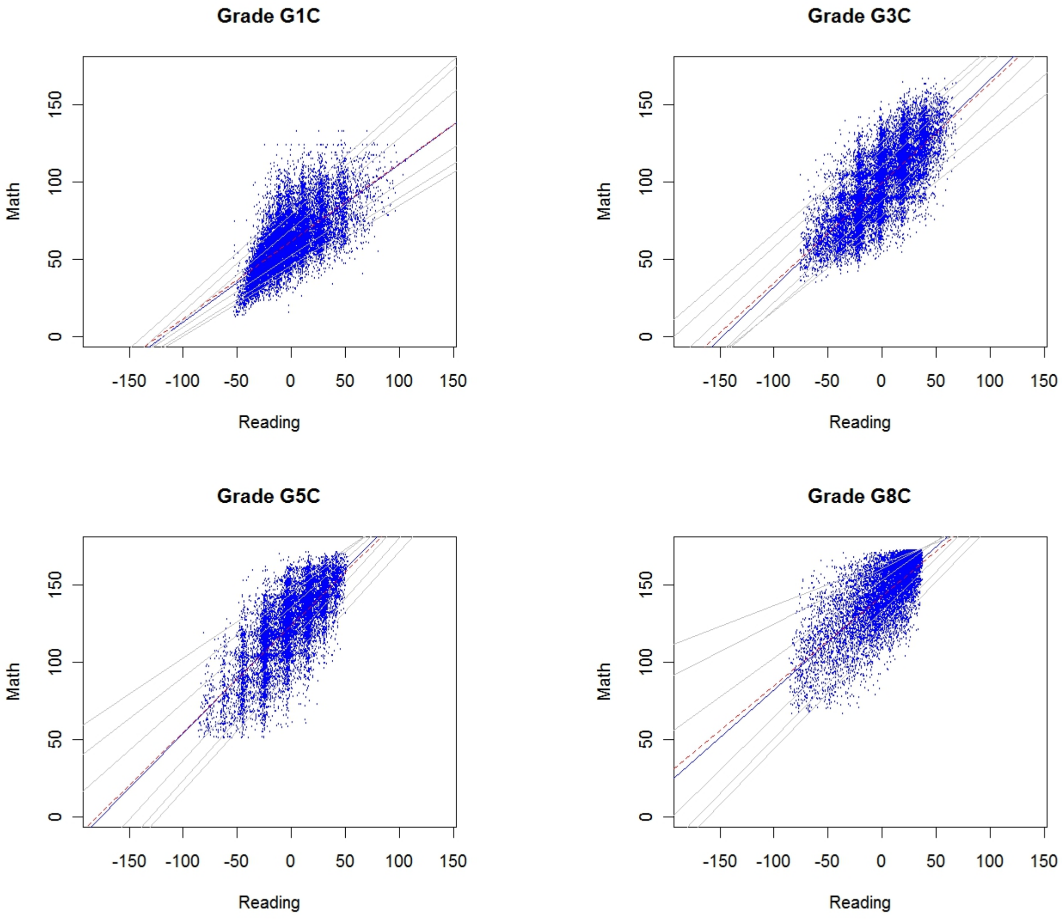
[13] Patterns related to grade 1 is very different from those in later grades with regard to many variables. This phenomenon is revealed throughout this chapter.

The process plot for the READING slope coefficient at grade 1 mimics the shape of the intercept, which means the language proficiency of students exerts a location and a scale shift of the math score distribution. The dispersion among high math ability students is larger than low math ability students. However, the variance becomes quite stable at grade 3 and is larger for low math ability students than high math ability students at grade 5 and 8. This phenomenon has already been demonstrated by the residual plots in Figure 3.1. It can also be easily seen through the relative position of the quantile regression lines as in Figure 4.3. While the lines diverge at high quanitles at grade 1, they converge at grade 5 and 8.

SES is strongly and positively related to math scores. At grade 1, the strength of the relationship increases as students' conditional math ability[14] increases. Tests of equivalence (Table 4.2) show that at later grades, the relationship is stable across quantiles. In other words, regardless of the conditional math ability, the influence of SES on the math scores stays the same in magnitude. The mean regression captures the relationship between SES and math scores as well as quantile regressions within these grades (Figure 4.1b).

GENDER has negative estimates for all quantiles and at all grades (Figure 4.1b). This is consistent with the majority of past research, that other things being equal, girls performed worse than boys on math assessment (Benbow & Stanley, 1980; Mau & Lynn, 2000). Test of equivalence (Table 4.2) shows that the gender gap varies between

[14] The terms "math ability" and "math scores" are used interchangeably because the scores are regarded to reflect students' current and end-product math ability regardless of other factors.



Grade G1C

Grade G3C

Grade G5C

Grade G8C

Figure 4.3 Simple Quantile Regression Plots

Note: The blue solid line is the mean regression line. The red dotted line is the median regression line. The gray lines represent the results corresponding to the 5th, 15th, 25th, 75th, 90th and 95th quantiles regression models.

The quantile models here are different from the complete regression model. Only READING is kept as the independent variable. Reading scores are all centered at the corresponding grade.

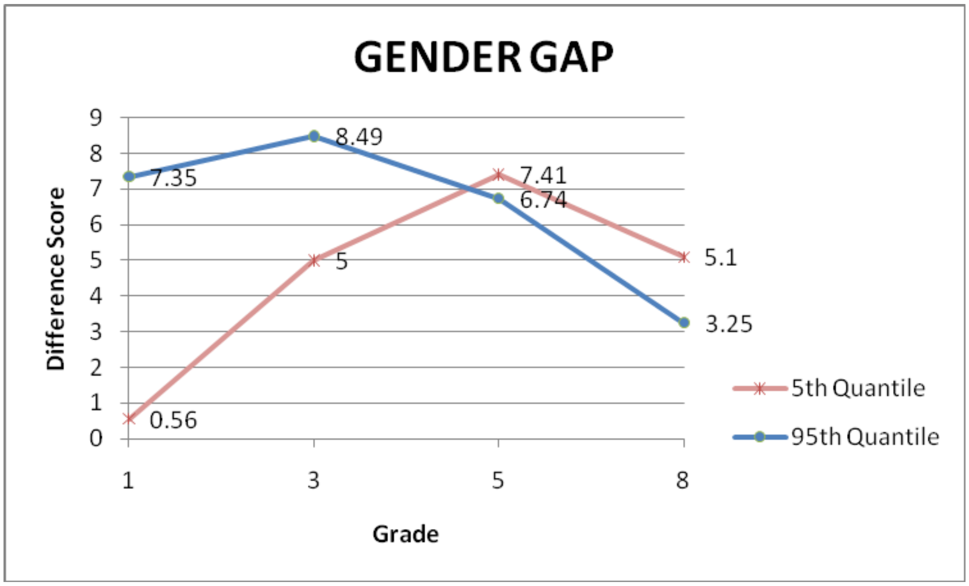


Figure 4.4 Plots of Score Difference between Boys and Girls

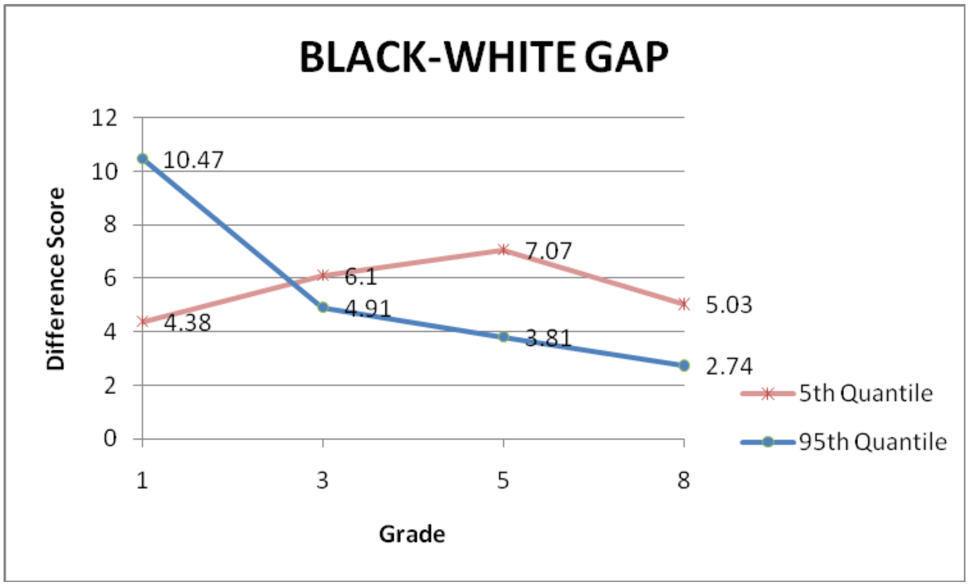


Figure 4.5 Plots of Score Difference between White and Black

quantiles for grade 1, 3 and 8. At grade 1, gender gap is bigger for high math ability students than for low math ability students. Or, the low math ability students are similarly low being a boy or a girl but the high math ability students differ more between girls and boys. For example, the score difference between a girl and a boy at the 5% percentile is less than 1 point at grade 1. The difference at the 95% percentile is more than 7 points (Figure 4.4). This trend seemed to change as students moved to higher grades. At grade 8, the gender gap is bigger for students with low math abilities (about 5 points at 5th math percentile) but smaller for high math abilities students (about 3 points at the 95th math percentile). Still, the results suggest that girls started school at a disadvantage compared to boys and this disadvantage continued as students grew in age.

Results for the subgroup BLACK are similar to GENDER. Consistent with past research, black students performed worse than their white counterparts (e.g. Tate, 1997) regardless of conditional math ability or grades (Figure 4.1d). Tests of equivalence (Table 4.2) show that the White-Black gap varies between quantiles only at grade 1 and 5. At grade 3 and 8, black students are homogeneously similar within group although still significantly worse than the white students. Figure 4.5 shows that at grade 1, the White-Black ethnicity gap is about 4 points at the 5th quantile and about 10 points at the 95th quantile. At grade 8, the gaps are about 5 points at the 5th quantile and 3 points at the 95th quantile. These results suggest that although black students started school at a disadvantage compared to white students, this disadvantage was more obvious for high math ability students at grade 1. The disadvantage observed in early grades for the black students remained as students moved into upper grades. However, schooling seems to have narrowed the gap within black students as they moved up the grades.

Central to the research questions is the ELL status impact. When READING is controlled for, neither Former-ELLs nor ELLs differ from non-ELLs at any grade. Although there are some sporadic significant results at some quantiles, they are not significant at most quantiles (Appendix D). Tests of equivalence (Table 4.2) showed that the coefficients are not significantly different from each other between quantiles. Based on this, the sporadic significant coefficients, such as for FORMER at the 25th quantile at grade 8, may be simply Type I errors. This is especially possible for the ELL effect since there are only 191 ELLs in total out of the complete sample. In brief, when READING is controlled for, there is no significant difference between the math scores of ELL, former ELL and non-ELL students. This result is different from many previous studies where gap is reported and the gap between these groups is considered increasing or decreasing (Abedi et al., 2005; Chang et al., 2009; Fry, 2007; Han, 2008). It seems that it is less the difference between these groups in math achievement scores and more the language proficiency that explains the variability. In the present data, since ELL status was decided based upon an oral language test, it also means that oral language proficiency is not a good indicator of students' math achievement. Rather, the math score difference between ELLs and Non-ELLs is likely the result of the reading skills required by the tests. Of course, not every reading subskill is necessary for math assessment, neither are the reading skills the same as the academic language proficiency for math learning. Still, the insignificance of ELL status in the presence of a reading score highlighted the language impact on math achievements for ELLs.

To further explore the language factor, models with and without READING are compared. For simplicity, only main effects are explored. The small difference in R^2 and pseudo- R^2 between models with and without interaction terms (Table 4.1) gives a foundation for this decision. The comparison results are summarized in Table 4.3.

Former ELLs do not differ from non-ELLs at grade 3, 5 and 8 regardless of the READING being controlled for or not. This suggests that once students have mastered certain level of oral language proficiency at early stage (grade 1), they can handle class work quite well to be comparable to their non-ELL peers later on. For grade 1, however, when READING is controlled for, there is no difference between Former ELLs and Non-ELLs; when READING is not controlled for, there is a difference between Former ELLs and Non-ELLs. This means language limitations still put the Former ELLs at disadvantage immediately after they are re-designated[15]. However, as the language proficiency of former ELLs continues to grow, by grade 3, these students can perform comparably well to their Non-ELL peers.

At all grades, when READING is controlled, math score difference between ELLs and non-ELLs are not statistically significant. When READING is not controlled, the difference is significant. This supports that math achievement gaps between ELLs and non-ELLs are mostly related to the language proficiency between these groups.

[15] The last round of language proficiency screening to decide ELL status stopped at Grade 1 for ECLS-K.

Table 4.3
Significance Comparison between Models with and without READING

		.05	.1	.25	.5	.75	.9	.95	MR
Grade 1	READING	+	+	+	+	+	+	+	+
	SES	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+
	GENDER	-	+	+	+	+	+	+	+
		+	+	-	+	+	+	+	+
	FORMER	-	-	-	-	-	-	+	-
		-	+	+	+	+	+	+	+
	ELL	-	-	-	-	-	-	-	-
		+	+	+	-	+	-	-	+
	BLACK	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+
	HISPANIC	-	+	+	+	+	+	+	+
		-	+	+	+	+	+	+	+
	ASIAN	+	+	+	+	+	-	-	+
		-	-	-	-	-	-	-	-
	ISOLATED	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+
	R^2	.30	.29	.28	.28	.29	.29	.26	.47
		.10	.11	.10	.10	.12	.12	.11	.20

		.05	.1	.25	.5	.75	.9	.95	MR
Grade 3	READING	+	+	+	+	+	+	+	+
	SES	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+
	GENDER	+	+	+	+	+	+	+	+
		-	+	+	+	+	+	+	+
	FORMER	-	-	-	-	-	-	-	-
		-	-	-	-	+	+	-	-
	ELL	-	-	-	-	-	-	-	-
		+	+	+	+	+	+	+	+
	BLACK	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+
	HISPANIC	-	+	+	-	-	-	-	+
		+	+	+	+	+	+	-	+
	ASIAN	-	-	-	-	+	+	-	-
		-	-	-	-	+	+	-	-
	ISOLATED	-	-	-	-	+	-	-	-
		+	+	+	+	+	+	+	+
	R^2	.32	.35	.38	.37	.35	.32	.30	.58
		.10	.13	.15	.14	.13	.11	.10	.25

Note: “+” indicates coefficient estimate that is significant at or below .05. “–“ indicates coefficient estimate that is insignificant at or below .05. The first row of the symbols are the results of the model with READING, the second row are the results of the model without READING. MR means mean regression results. The gray areas are of interest for the original comparison purpose. The yellow areas represent variables that changed significance levels between models.

Table 4.3 Continued

		.05	.1	.25	.5	.75	.9	.95	MR			.05	.1	.25	.5	.75	.9	.95	MR
Grade 5	READING	+	+	+	+	+	+	+	+	Grade 8	READING	+	+	+	+	+	+	+	+
	SES	+	+	+	+	+	+	+	+		SES	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	+
	GENDER	+	+	+	+	+	+	+	+		GENDER	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+			-	-	+	+	+	+	+	+
	FORMER	-	-	+	-	-	-	-	-		FORMER	-	-	+	-	-	-	-	-
		-	-	-	-	-	-	-	-			-	+	-	-	-	-	-	-
	ELL	-	-	-	-	-	-	-	-		ELL	-	-	-	-	-	-	-	-
		+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	+
	BLACK	+	+	+	+	+	+	+	+		BLACK	+	+	+	+	+	+	+	+
		+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	+
	HISPANIC	-	-	-	-	+	-	-	+		HISPANIC	-	-	-	-	-	-	-	-
		+	-	+	+	+	+	-	+			+	-	+	+	+	+	+	+
	ASIAN	+	+	+	+	+	+	+	+		ASIAN	-	+	+	+	+	+	+	+
		-	+	+	+	+	+	+	+			-	+	+	+	+	+	+	+
	ISOLATED	-	-	-	+	+	-	-	-		ISOLATED	-	-	+	-	-	-	-	+
		+	+	+	+	+	+	+	+			+	+	+	+	+	-	-	+
	R^2	.39	.40	.40	.36	.32	.27	.23	.58		R^2	.42	.41	.39	.35	.29	.21	.16	.57
		.15	.17	.17	.15	.13	.10	.08	.26			.17	.17	.17	.15	.11	.07	.05	.25

Note: "+" indicates coefficient estimate that is significant at or below .05. "–" indicates coefficient estimate that is insignificant at or below .05. The first row of the symbols are the results of the model with READING, the second row are the results of the model without READING. MR means mean regression results. The gray areas are of interest for the original comparison purpose. The yellow areas represent variables that changed significance levels between models.

In terms of the ethnicity group difference, at grades 3, 5 and 8, when language proficiency is controlled for, the HISPANIC-WHITE and ISOLATED-WHITE gaps are not significant. When language proficiency is not controlled for, they are significant at most conditional math ability levels. Combined with results from Figure 4.1d and 4.1e, this means that the low math achievement of the Hispanics and the isolated students is related to their language proficiency. The change in significance for the ISOLATED seem to suggest that although the geographically isolated students are rarely considered ELLs by common definitions, their language proficiency actually have put them at disadvantage in their performance in math assessment. In contrary to these two groups, Asians, however, do not seem to differ with READING controlled for or not.

All these observations described so far are consistent at grade 3, 5 and 8. Patterns at grade 1 are opposite. Collinearity diagnosis reveals that Asian and Hispanics are correlated to Former ELL or ELL status. This might explain the unique pattern for the Asians. However, the variance inflation factor (VIF) is low and is less than 1.8 even at the extreme. For this reason, these two variables are kept in the model together with ELL status. However, there are only 191 ELLs out of the complete sample, which explains the width of the confidence interval around the estimates for ELLs (Figure 4.1c). The confidence interval for ELLs is not even displayed in Figure 4.1c because it is far beyond the vertical range used for the graphs. Like for many other effects described before, grade 1 stands out different from all later grades. Student performance at this grade seems quite varied and hard to interpret.

Test of equivalence (Table 4.2) shows that the estimates for HISPANIC, ASIAN and ISOLATED are similar across quantiles. This means, mean regression gives as much information as quantile regressions for these three variables. Based on Table 4.2 and Figure 4.1, it can be summarized that Hispanic students performed worse than their White counterparts at grade 1 and 3 but caught up at grade 5 and 8. Students from the geographically isolated areas started worse than their White counterpart at grade 1 but caught up by grade 3. Asian students, uniquely, performed worse than their White peers at grade 1 but caught up by grade 3. More impressively, they outperformed their White peers at both grade 5 and 8.

Traditionally it is regarded that main effects should be interpreted with interaction terms if the latter are significant. Thus so far, the main effects are discussed in terms of general trend. As to be seen soon, interaction terms moderated the magnitude of the main effects but did not change the direction of the main effects. However, since the grade level models are not the statistically best-fitting model at each grade, interaction terms are not discussed here. More specific interpretation is presented in a different way through the longitudinal model.

Longitudinal Model Result

Model Building

Model building is an important step to reach a statistically best-fitting model. Results can be interpreted with more confidence when the statistics are based on the best-fitting model. The model building process can also provide better guidance for readers to understand and evaluate the current research for their own benefit. For this purpose, the process is described in great detail in this section.

The longitudinal model was built step by step from a full model (Model 1) with all relevant 2-way and 3-way interactions. Models were compared and evaluated at each step by looking at the R^2 and significance pattern of each coefficient. High-way interactions are examined first and the main effects last. However, to avoid fishing and to be efficient, a systematic and abbreviated strategy is used. Whenever 3-way interactions are under examination, all 2-way interactions are kept in the reduced model even if they are insignificant. After decisions about 3-way interactions are finalized through model comparison, 2-way interactions are then studied. Whenever 2-way interactions are under examination, all main effects are kept in the reduced model even if they are insignificant.

There are two reasons a term is removed. First, it is consistently insignificant at an alpha level of .05 at all quantiles in all the models up to the one under examination. An example of this is the Reading*ELL interaction which is consistently insignificant at the .05 level in both Model 1 and Model 2. Thus it has been removed since Model 3. Second, a variable is removed if the significance level jumps between quantiles and/or models although it is significant at or below alpha level of .05 at some quantiles. An example of this is the Reading*Hispanic*Time. This term is significant (p=.03) at the 10th quantile in Model 1 but is not significant at any quantile in Model 2. It is thus removed starting from Model 3. Once a term is removed, it is not considered in the following models again. In this sense, this procedure falls into the backward elimination family. It is an improved procedure because the decision to eliminate a term is postponed to the next model results rather than on the immediate results. This is necessary because for quantile regression modeling, there are usually several models evaluated at the same time. However, least absolute deviance estimators are unstable. A tiny change in the data can lead to relatively big changes in the fitted plane (Good & Hardin, 2006, pp.165). To be conservative, decisions are better made basing on consistent results from more than one model.

The model comparison results are represented in Figure 4.6. It revealed the significance level change of all the variables kept until the final reduced model. The color of the bars for each variable indicates the significance level. Dark gray means the coefficient is significant at or below .01 and medium gray means significant at or below .05. Light gray indicates significance level higher than .05 but lower than .1. This last category is included just for information but is not discussed in the current study. The vertical scale refers to the specific quantile regression the result is based on.

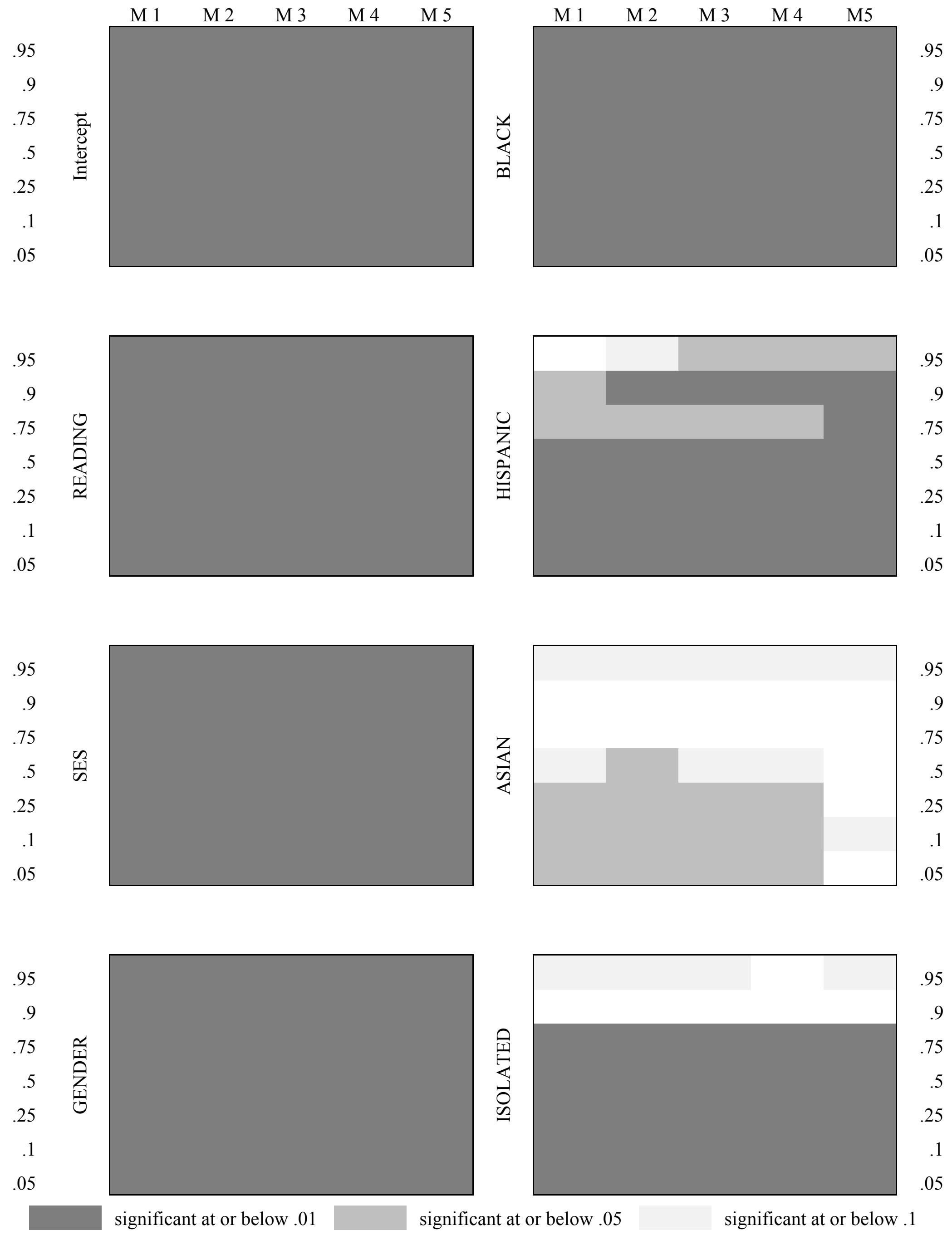


Figure 4.6a Main Effects

Note: M1 is the most complicated model with all the terms of original interest. M5 is the final reduced model with mostly significant terms only. Vertical scale indicates the corresponding quantile regression within each model.

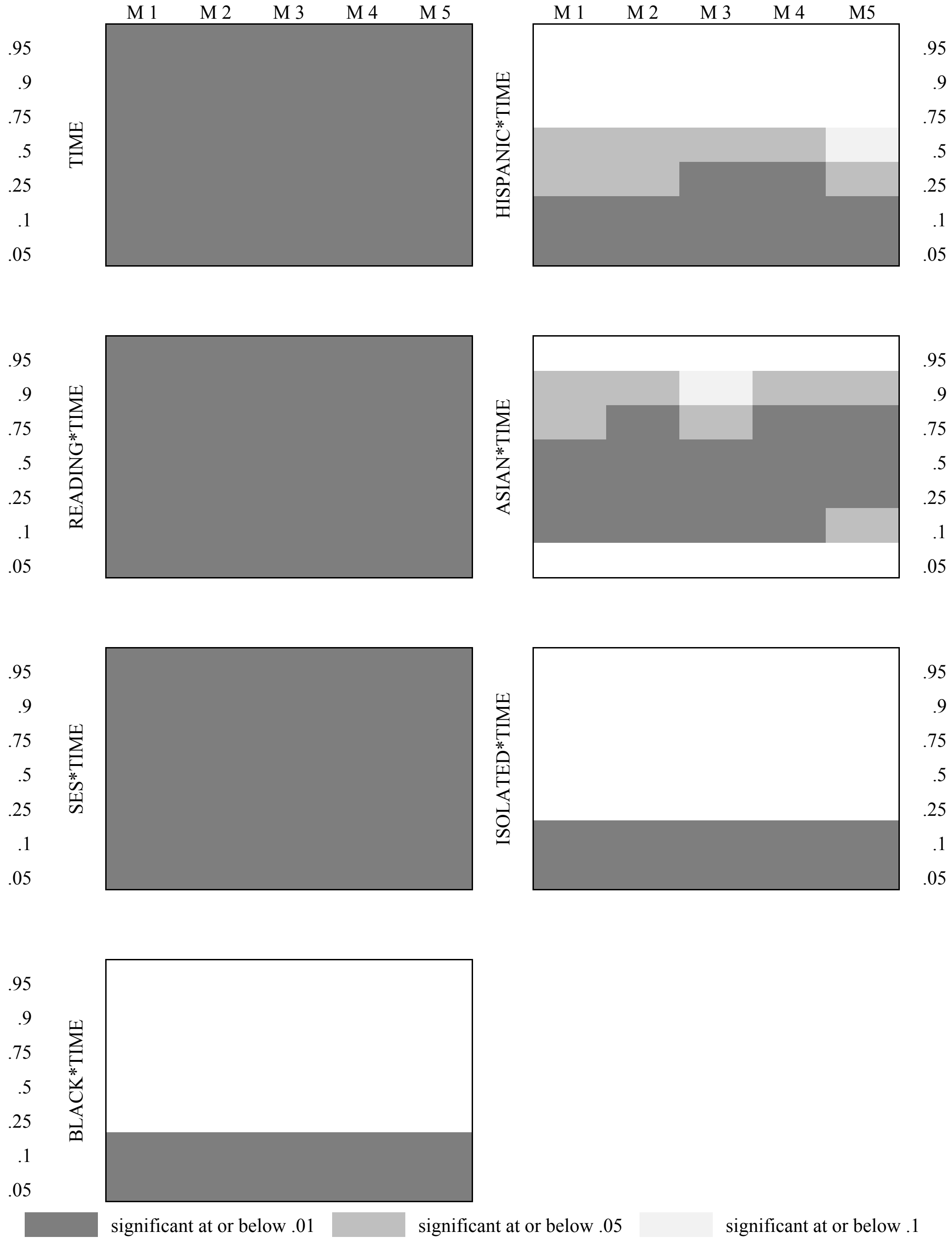


Figure 4.6b Two-way Interaction

Note: M1 is the most complicated model with all the terms of original interest. M5 is the final reduced model with mostly significant terms only. Vertical scale indicates the corresponding quantile regression within each model.

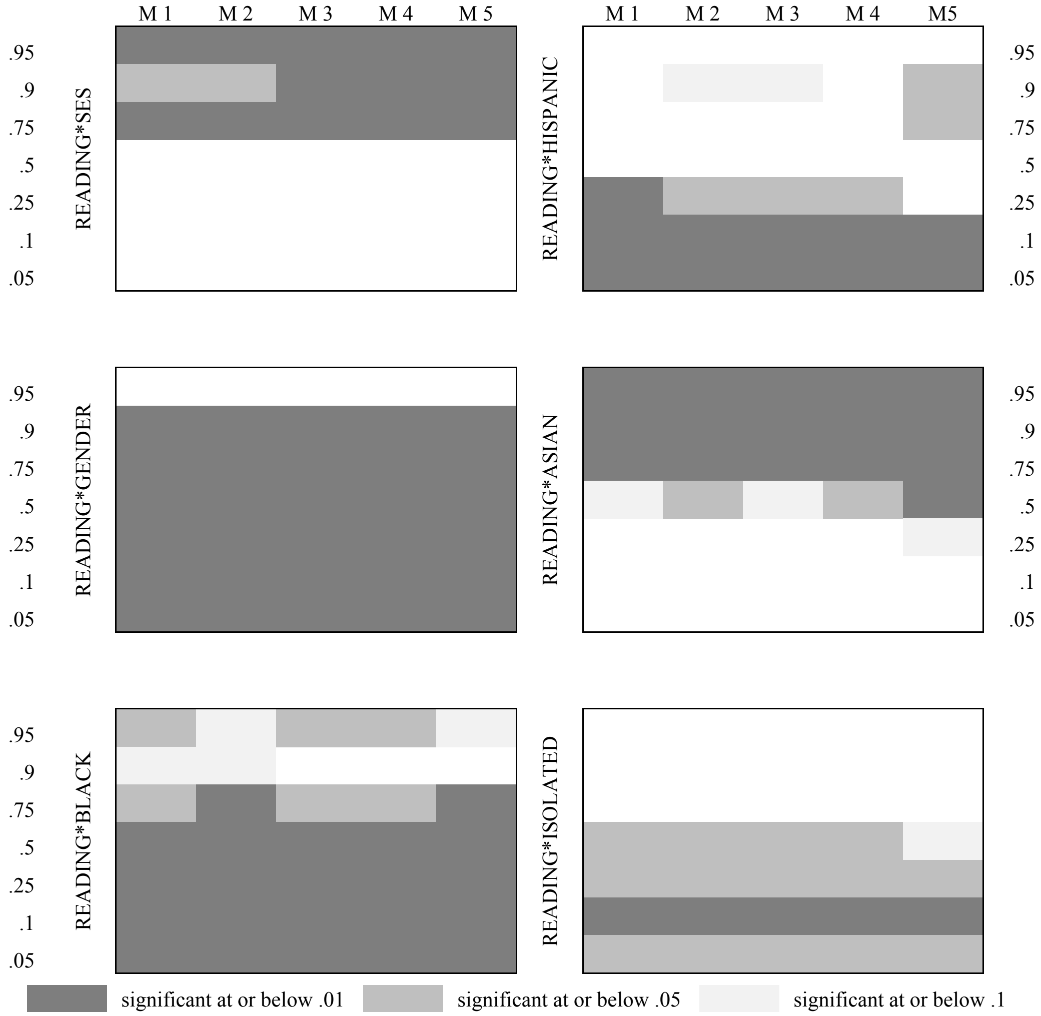


Figure 4.6c Two-way Interaction Continued

Note: M1 is the most complicated model with all the terms of original interest. M5 is the final reduced model with mostly significant terms only. Vertical scale indicates the corresponding quantile regression within each model.

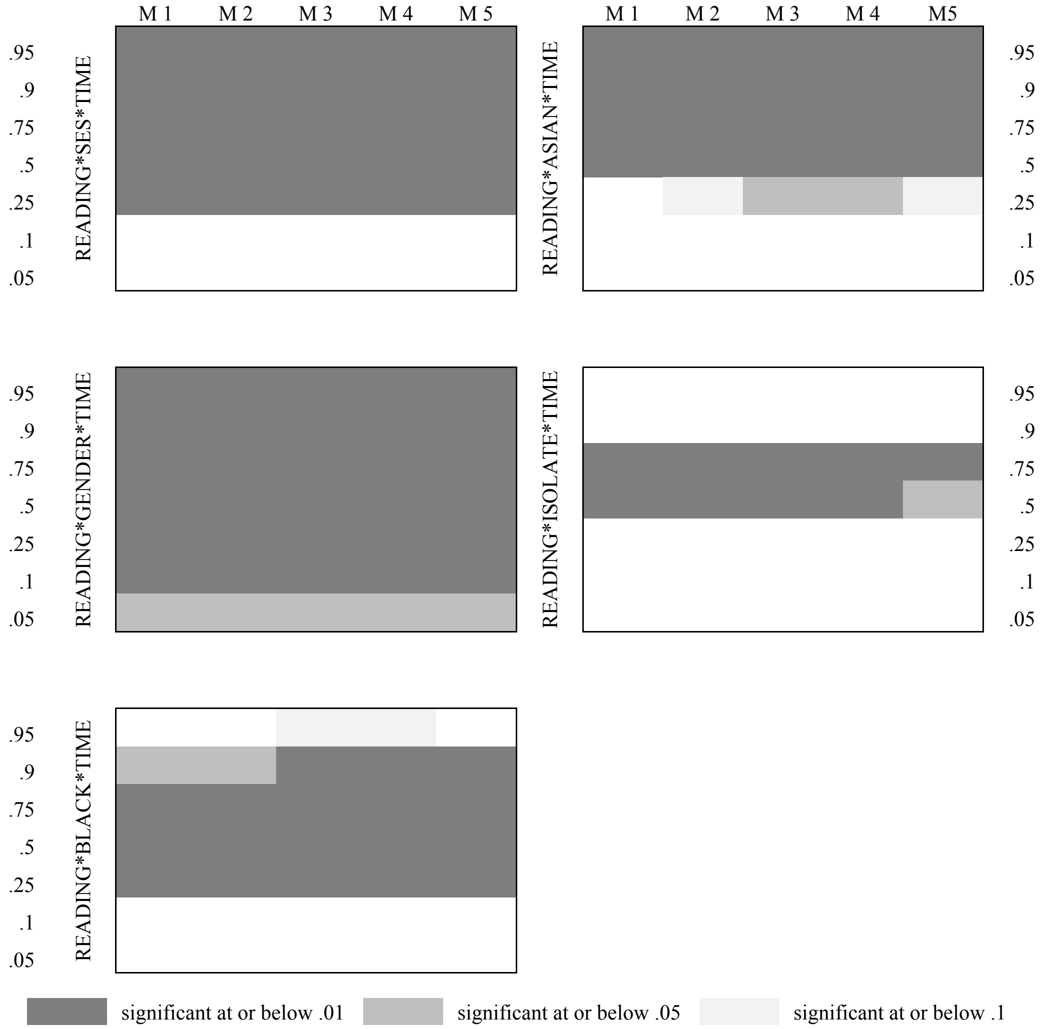


Figure 4.6d Three-way Interaction

Note: M1 is the most complicated model with all the terms of original interest. M5 is the final reduced model with mostly significant terms only. Vertical scale indicates the corresponding quantile regression within each model.

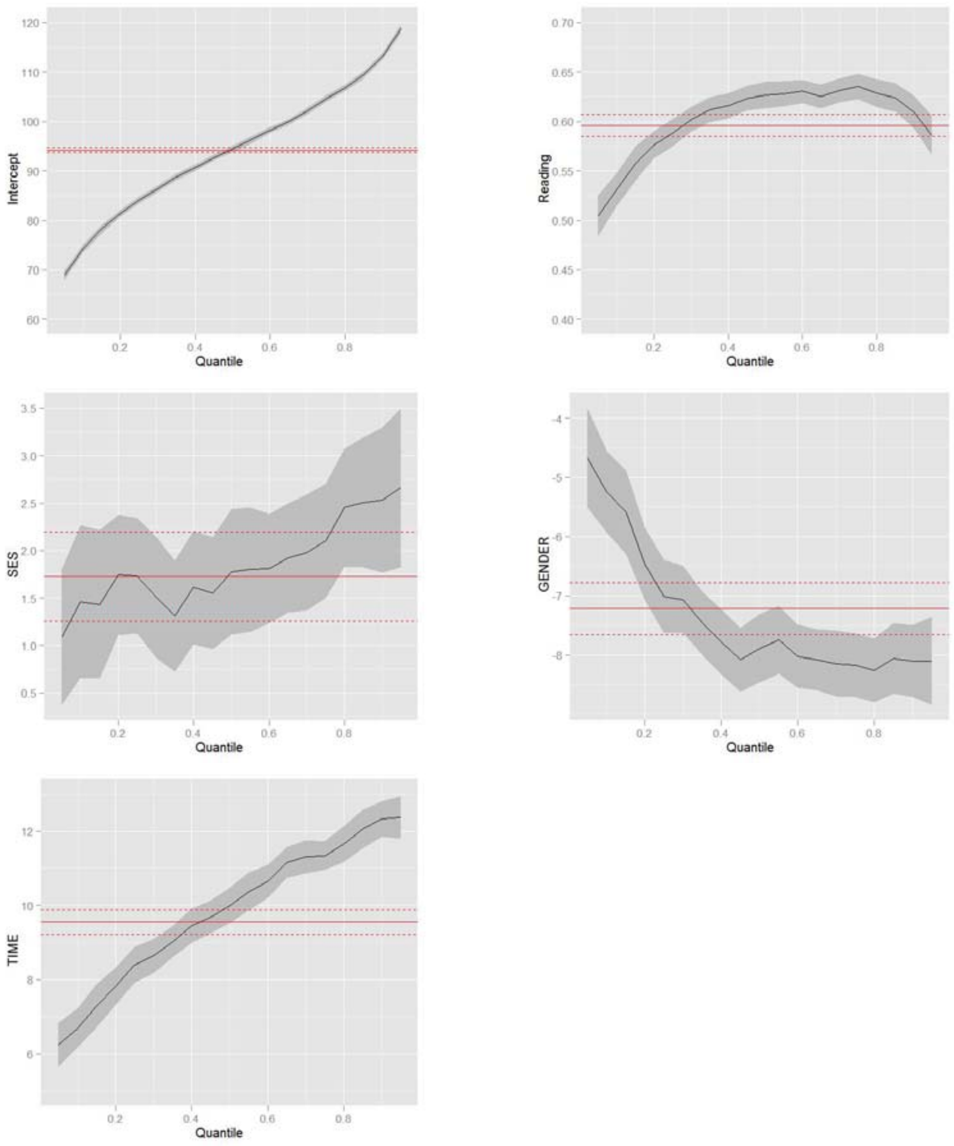


Figure 4.7a Main Effects

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

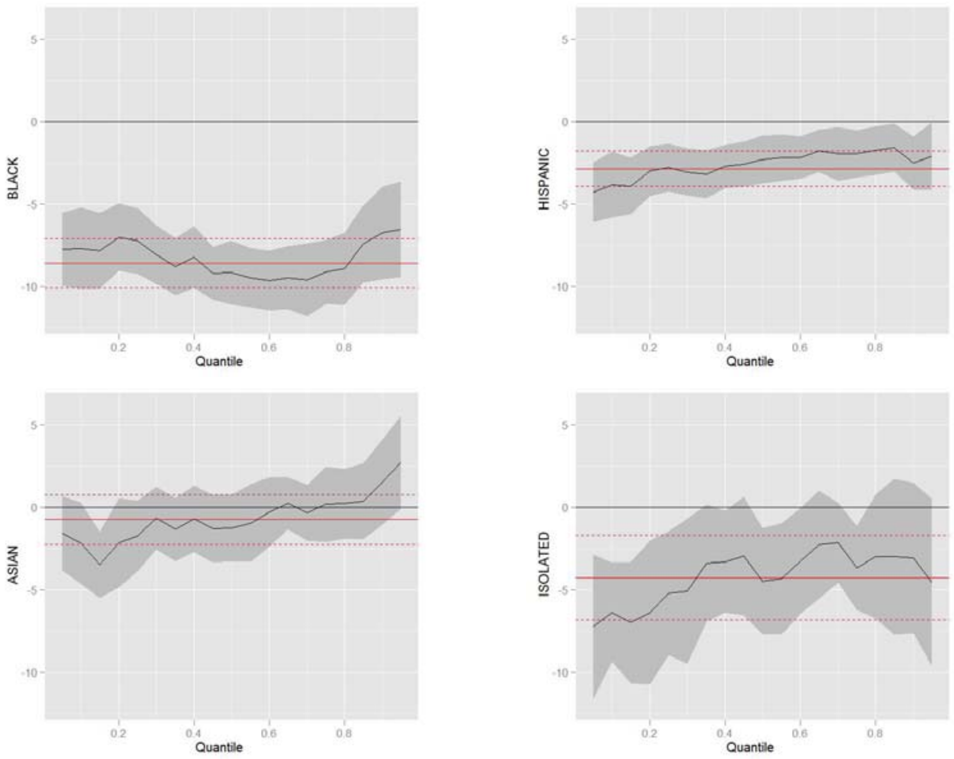


Figure 4.7b Main Effects: Race-Ethnicity

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

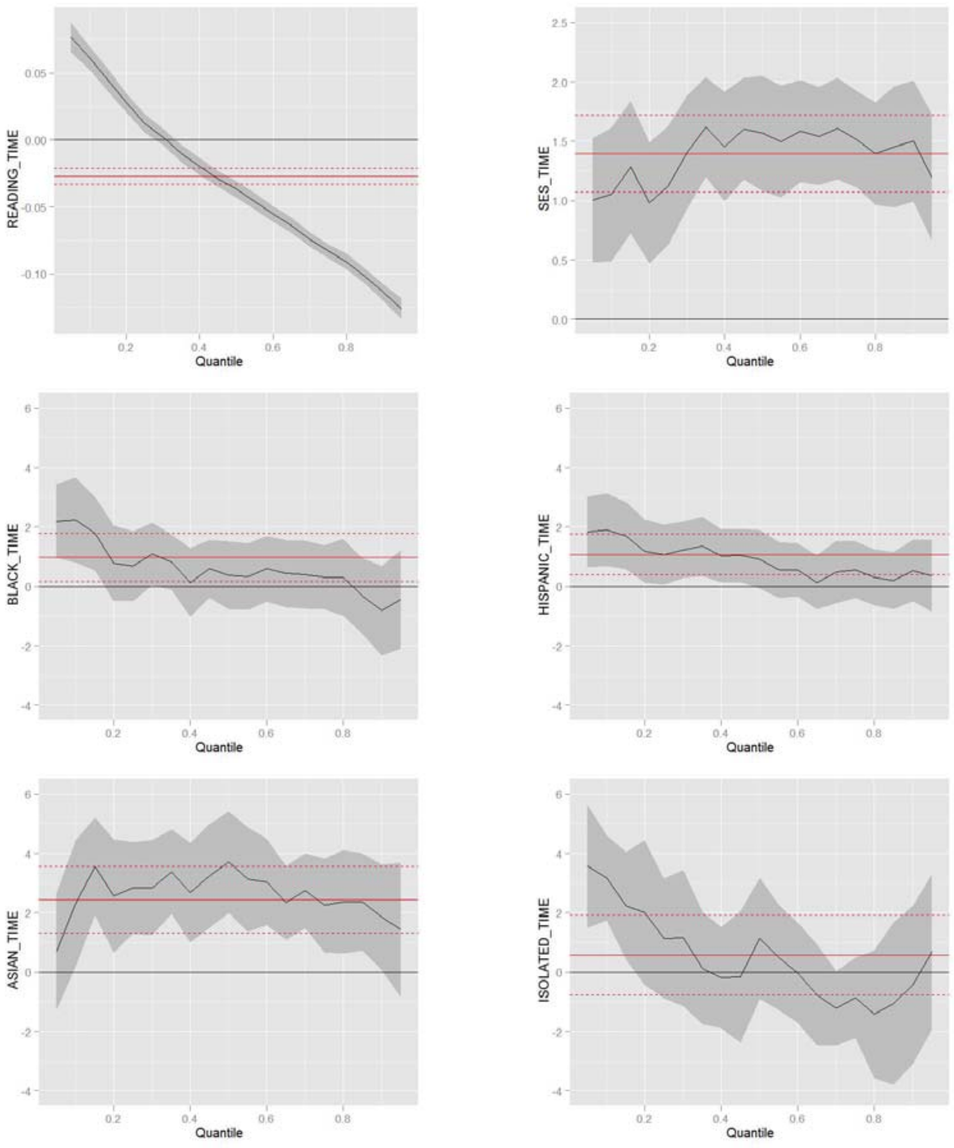


Figure 4.7c Two-way Interaction

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

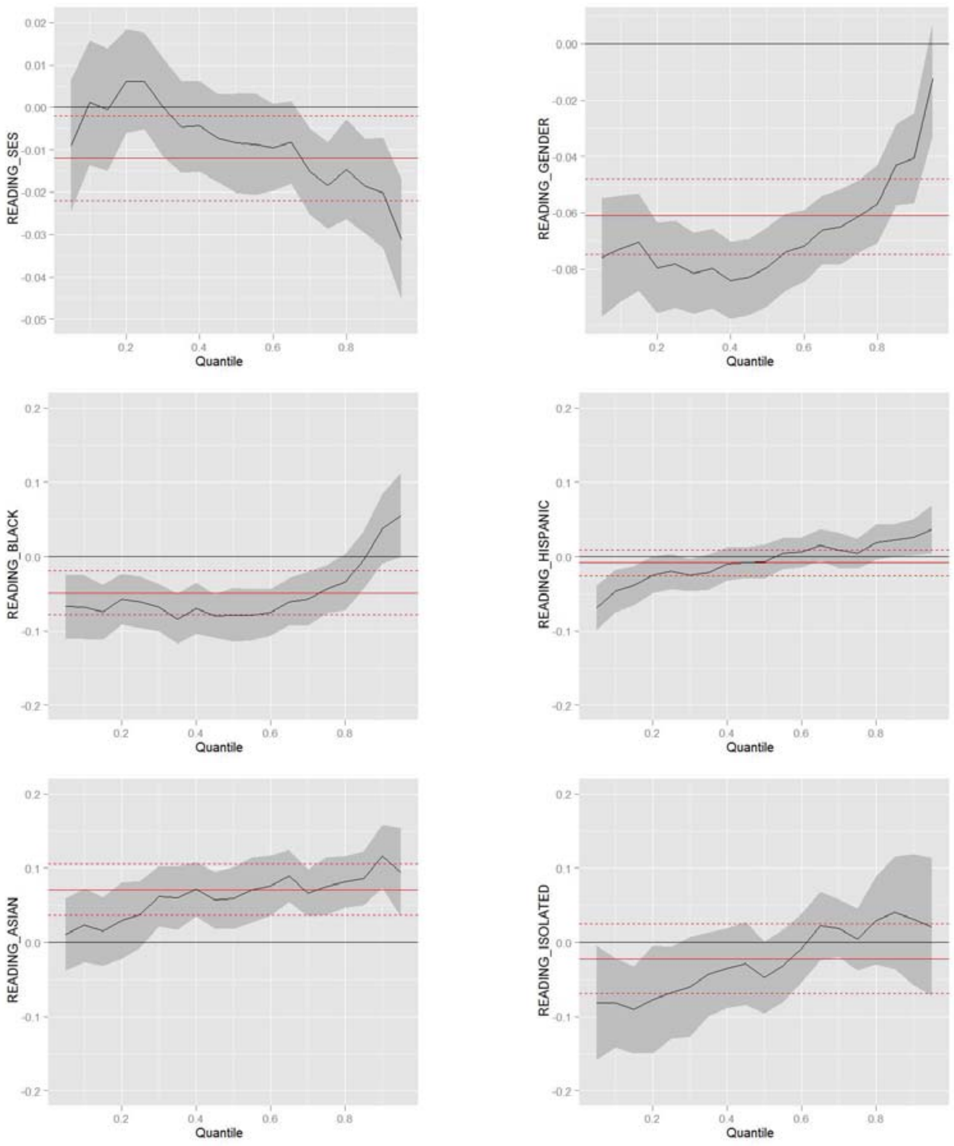


Figure 4.7d Two-way Interaction Continued

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

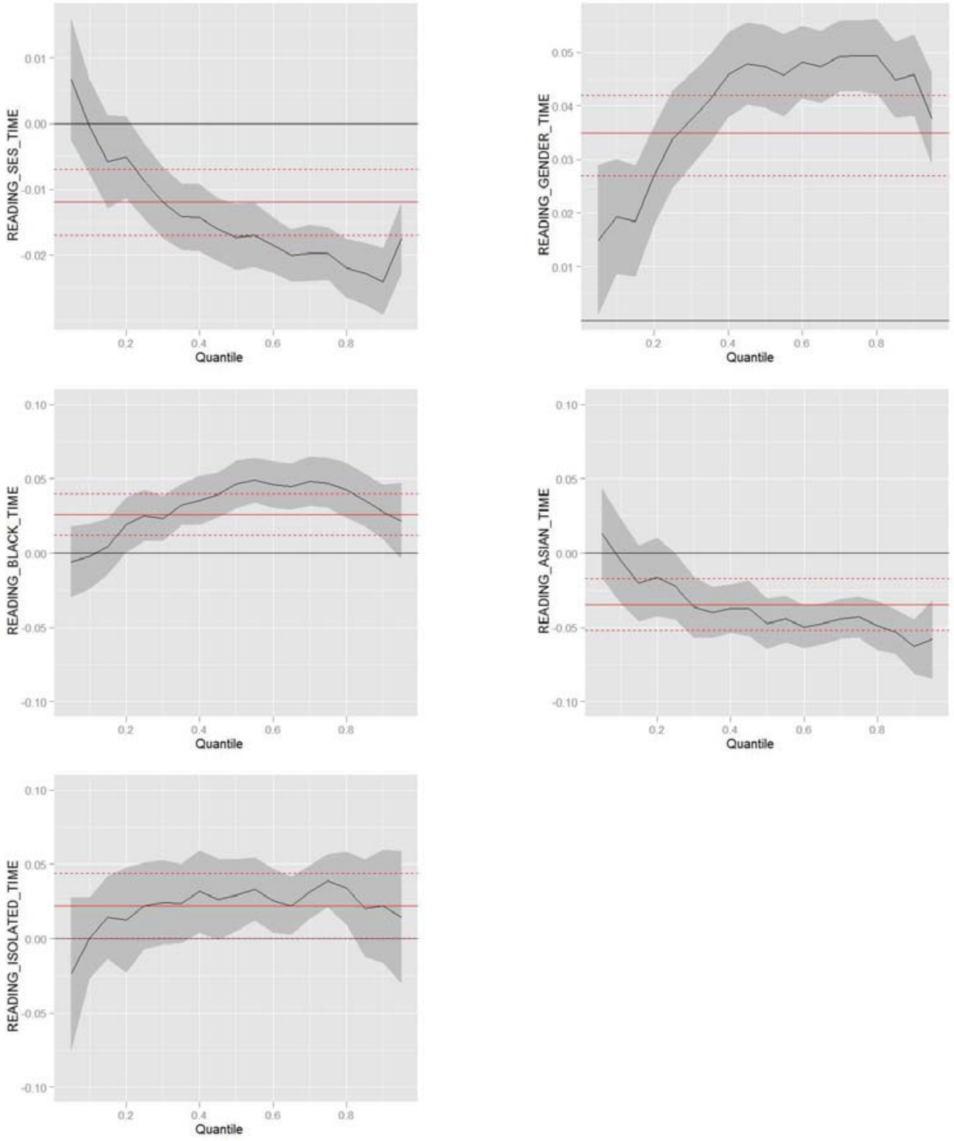


Figure 4.7e Three-way Interaction

The x axis represents the quantile range from 0 to 1. The extreme values below .05 and .95 are not graphed because they tend to be unstable as extremes. The y axis represents the intercept or slope coefficient corresponding to each specific variable at each quantile.
The black line with shade area represents the quantile regression results with 95% confidence interval.
The red solid line is the mean regression results with dotted lines representing the 95% confidence interval.
A black horizontal reference line at y=0 is also graphed when possible for hypothesis testing.

Table 4.4a
Longitudinal Model Results : Main Effects

	.05	.10	.25	.50	.75	.90	.95	MR
Intercept	68.88	74.27	84.07	94.44	104.61	113.13	118.73	94.23
	(.49)	(.41)	(.35)	(.33)	(.34)	(.37)	(.47)	(.26)
	***	***	***	***	***	***	***	***
READING	.50	.53	.59	.63	.64	.61	.59	.60
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
	***	***	***	***	***	***	***	***
SES	1.09	1.46	1.74	1.77	2.10	2.53	2.66	1.73
	(.37)	(.41)	(.31)	(.34)	(.31)	(.39)	(.43)	(.24)
	***	***	***	***	***	***	***	***
GENDER	-4.67	-5.24	-7.01	-7.89	-8.17	-8.10	-8.10	-7.22
	(.43)	(.35)	(.32)	(.3)	(.27)	(.31)	(.38)	(.22)
	***	***	***	***	***	***	***	***
BLACK	-7.75	-7.71	-7.24	-9.18	-9.13	-6.76	-6.54	-8.61
	(1.13)	(1.26)	(1.02)	(.99)	(.98)	(1.43)	(1.47)	(.76)
	***	***	***	***	***	***	***	***
HISPANIC	-4.29	-3.82	-2.79	-2.33	-1.97	-2.53	-2.12	-2.86
	(.92)	(1.02)	(.75)	(.74)	(.72)	(.82)	(1.05)	(.55)
	***	***	***	***	***	***	**	***
ASIAN	-1.58	-2.18	-1.76	-1.26	.17	1.53	2.71	-.74
	(1.15)	(1.25)	(1.09)	(1.04)	(1.16)	(1.32)	(1.44)	(.77)
		*					*	
ISOLATED	-7.23	-6.37	-5.20	-4.49	-3.67	-3.07	-4.58	-4.27
	(2.22)	(1.53)	(1.92)	(1.64)	(1.29)	(2.33)	(2.59)	(1.3)
	***	***	***	***	***		*	***
TIME	6.25	6.70	8.40	9.99	11.34	12.33	12.37	9.55
	(.3)	(.27)	(.25)	(.24)	(.2)	(.24)	(.29)	(.17)
	***	***	***	***	***	***	***	***

Note:
MR refers to mean regression. The first rows are estimates, second rows are standard errors. Third rows are significance level.
*** indicates significance level at or below .01
** indicates significance level at or below .05
*indicates significance level at or below .1

Table 4.4b
Longitudinal Model Results : Two-way Interaction

	.05	.1	.25	.5	.75	.90	.95	MR
READING*TIME	.08	.06	.01	-.04	-.08	-.11	-.13	-.03
	(.01)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
	***	***	***	***	***	***	***	***
SES*TIME	1.00	1.05	1.12	1.57	1.52	1.50	1.19	1.40
	(.27)	(.29)	(.26)	(.25)	(.21)	(.26)	(.27)	(.17)
	***	***	***	***	***	***	***	***
BLACK*TIME	2.20	2.24	.68	.38	.31	-.82	-.44	.98
	(.63)	(.74)	(.6)	(.58)	(.55)	(.77)	(.85)	(.41)
	***	***						**
HISPANIC*TIME	1.84	1.90	1.06	.91	.56	.54	.35	1.08
	(.61)	(.62)	(.52)	(.5)	(.48)	(.53)	(.61)	(.34)
	***	***	**	*				***
ASIAN*TIME	.68	2.33	2.83	3.70	2.25	1.85	1.42	2.44
	(1)	(1.09)	(.79)	(.87)	(.81)	(.91)	(1.16)	(.57)
		**	***	***	***	**		***
ISOLATED*TIME	3.57	3.17	1.14	1.14	-.87	-.43	.69	.58
	(1.06)	(.73)	(1.03)	(1.04)	(.69)	(1.36)	(1.34)	(.68)
	***	***						
READING*SES	-.01	.00	.01	-.01	-.02	-.02	-.03	-.01
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
					***	***	***	**
READING*GENDER	-.08	-.07	-.08	-.08	-.06	-.04	-.01	-.06
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
	***	***	***	***	***	***		***
READING*BLACK	-.07	-.07	-.06	-.08	-.04	.04	.06	-.05
	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.03)	(.02)
	***	***	***	***	***		*	***
READING*HISPANIC	-.07	-.05	-.02	-.01	.00	.03	.04	-.01
	(.02)	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)	(.01)
	***	***				**	**	
READING*ASIAN	.01	.02	.04	.06	.08	.12	.09	.07
	(.02)	(.03)	(.02)	(.02)	(.02)	(.02)	(.03)	(.02)
			*	***	***	***	***	***
READING*ISOLATED	-.08	-.08	-.07	-.05	.00	.03	.02	-.02
	(.04)	(.03)	(.03)	(.02)	(.02)	(.04)	(.05)	(.02)
	**	***	**	*				

Note:
MR refers to mean regression. The first rows are estimates, second rows are standard errors. Third rows are significance level.
*** indicates significance level at or below .01
** indicates significance level at or below .05
*indicates significance level at or below .1

Table 4.4c
Longitudinal Model Results : Three-way Interaction

	.05	.1	.25	.5	.75	.90	.95	MR
	.01	.00	-.01	-.02	-.02	-.02	-.02	-.01
READING*SES*TIME	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
			***	***	***	***	***	***
	.01	.02	.03	.05	.05	.05	.04	.03
READING*GENDER*TIME	(.01)	(.01)	(0)	(0)	(0)	(0)	(0)	(0)
	**	***	***	***	***	***	***	***
	-.01	.00	.03	.05	.05	.03	.02	.03
READING*BLACK*TIME	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
			***	***	***	***		***
	.01	.00	-.02	-.05	-.04	-.06	-.06	-.03
READING*ASIAN*TIME	(.02)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
			*	***	***	***	***	***
	-.02	.00	.02	.03	.04	.02	.01	.02
READING*ISOLATE*TIME	(.03)	(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.01)
				**	***			

Note:
MR refers to mean regression. The first rows are estimates, second rows are standard errors. Third rows are significance level.
*** indicates significance level at or below .01
** indicates significance level at or below .05
*indicates significance level at or below .1

Results

It is clear that the pattern of the significance of the remaining variables is quite consistent both between models and within quantiles. The stability of the results between models gives confidence that there is no serious collinearity issue. Stability within quantiles guides the necessity to interpret the quantile regression results that stand out consistently. For example, READING is consistently significant. This gives strong support to the relationship between READING and math achievement. The Asian*TIME interaction term changes in significance level at some quantiles between models, but the coefficients for quantiles between .1 to .9 are always significant. In this case, the interpretation of the effect within this region of math score distribution is guaranteed.

Interpretation of the modeling results are also facilitated by two other sources of information from Figure 4.7 and Table 4.4. Figure 4.7 plots the final quantile regression estimates based on 19 quantiles and Table 4.4 summarizes the statistical results for the seven quantiles at the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentile for the same models.

Unlike the grade level models in the previous section, READING is centered around the grand mean across four grades here. The grades are coded as "TIME" with values of 0, 1, 2 and 3. The intercept thus refers to the math score of a White, male, non-ELL student with average SES background and average across-grade language proficiency at grade 1. Similar to Figure 4.1 and 4.2, x-axis marks the quantiles and the y-axis the specific intercept or slope estimate. The black solid line with shaded bands represents the quantile regression results and the red lines the mean regression results. When possible, a black vertical line at y=0 is also plotted for null hypothesis.

Results are similar as in the separate grade level models. However, since all the data is put into the same model, trends across time can be interpreted with more confidence. Math and reading scores are both vertically scaled using 3PL item response theory (Najarian, Pollack & Sorongon, 2009). The assumption here is that the vertical scale is valid. This assumption, also true for regular mean regression, is less of concern in quantile regression modeling since quantile regression uses order statistics and is free from normality assumption. As long as the relative standing of the students' ability maintains, the pattern and trend of the relationship between language and math remains.

Many main effects are consistently and statistically significant. Most two-way and three-way interactions are also significant. However, the statistical significance may be less meaningful when the estimates are in the hundredth of one point. This is an observation in the current study for many two-way and three-way terms. As Good and Hardin (2006) discussed, higher-order interaction terms can neither be given practical interpretation nor "have real meaning" (p.67). They can be examples of Type I errors and should not be over-interpreted. With these cautions in mind, rather than giving routine but meaningless sentences such as "GENDER modifies the interaction between READING and TIME by …", three –way interactions will be presented in graphics to understand the meaning in a more concrete way.

Just as in previous grade level models, READING is strongly and positively related to math achievement at all grades. The magnitude of the relationship, however, follows a concave down shape (Figure 4.7a). Overall, the strongest relationship is at the center bulk of the math score distribution. READING*TIME is positive at lower quantiles and negative at higher quantiles and the absolute value tends to be larger at higher quantiles than at lower quantiles. This could mean two things. First, the relationship between language and math achievement is not the same across time. Second, while the relationship is getting stronger for lower math ability students, it is getting weaker for higher math ability students over time. These interpretations are best seen in Figure 4.8 where the specific coefficients at each quantile and grade are calculated and plotted.

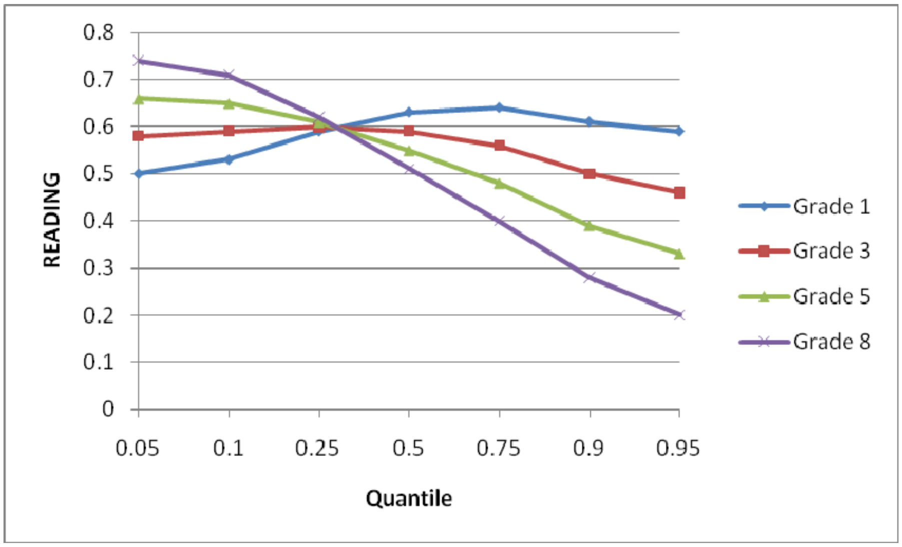


Figure 4.8 READING Effect at Each Grade
Note: Grade level results are calculated from the estimates in the longitudinal model where READING is centered around the grand mean of the vertical scale.

Just like the grade level model results, the relationship between READING and MATH is unique at grade 1. While the strength of the relationship keeps on increasing as math ability increases until the 75th percentile of conditional math ability at this grade, it evidently decreases at all quantiles for all other grades. The climbing relationship at quantile below .75 at grade 1 may be due to the fact that more and more math concepts start to rely on language skills to be expressed and understood (Ausubel & Robinson, 1969). Students at this grade are thus more sensitive to the language factor as their math skills expand. However, once some language proficiency threshold is reached, the language impact starts to decrease. The declining relationship at high math ability and at later grades may indicate that only some reading skills--- closer to academic language that is specific to the math content --- continue to affect math performance in tests. In the end, math and reading are two distinct subjects; the overlap between them (vocabulary, syntax) may be diminishing fast. A reasonable hypothesis is that students' high math ability can *compensate* for some possible deficiency in language skills in test performance. This offset may have happened because the linguistic threshold (Burns et al., 1983; Cummins, 1979b) is reached that minimizes the role language plays in math learning. In the same vein, high math ability have counterbalanced the language requirement in math assessment. Of course, other significant interaction terms involving READING moderate this relationship. The specific READING influence for each gender or ethnicity is discussed separately below.

SES is strongly and positively related to math achievement. In addition, SES*TIME is positive, meaning that the relationship between SES and math continues to increase as students grow (Figure 4.9). In all, high SES keeps on benefiting students at all math ability levels.

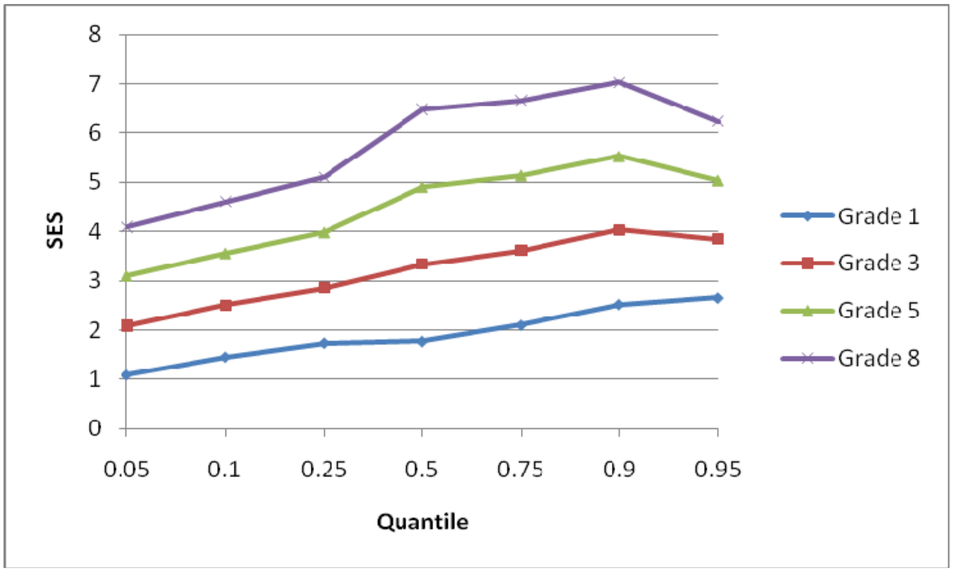


Figure 4.9 SES Effect at Each Grade

GENDER effect is the same as previously discussed. GENDER*TIME is not significant and has been removed from the final model. In summary, girls started first grade with lower math scores than boys and the gap continued throughout later grades. Low math ability students are similar whether a boy or a girl. However, for other ability levels, girls performed much worse than boys in math assessments (Table 4.4a).

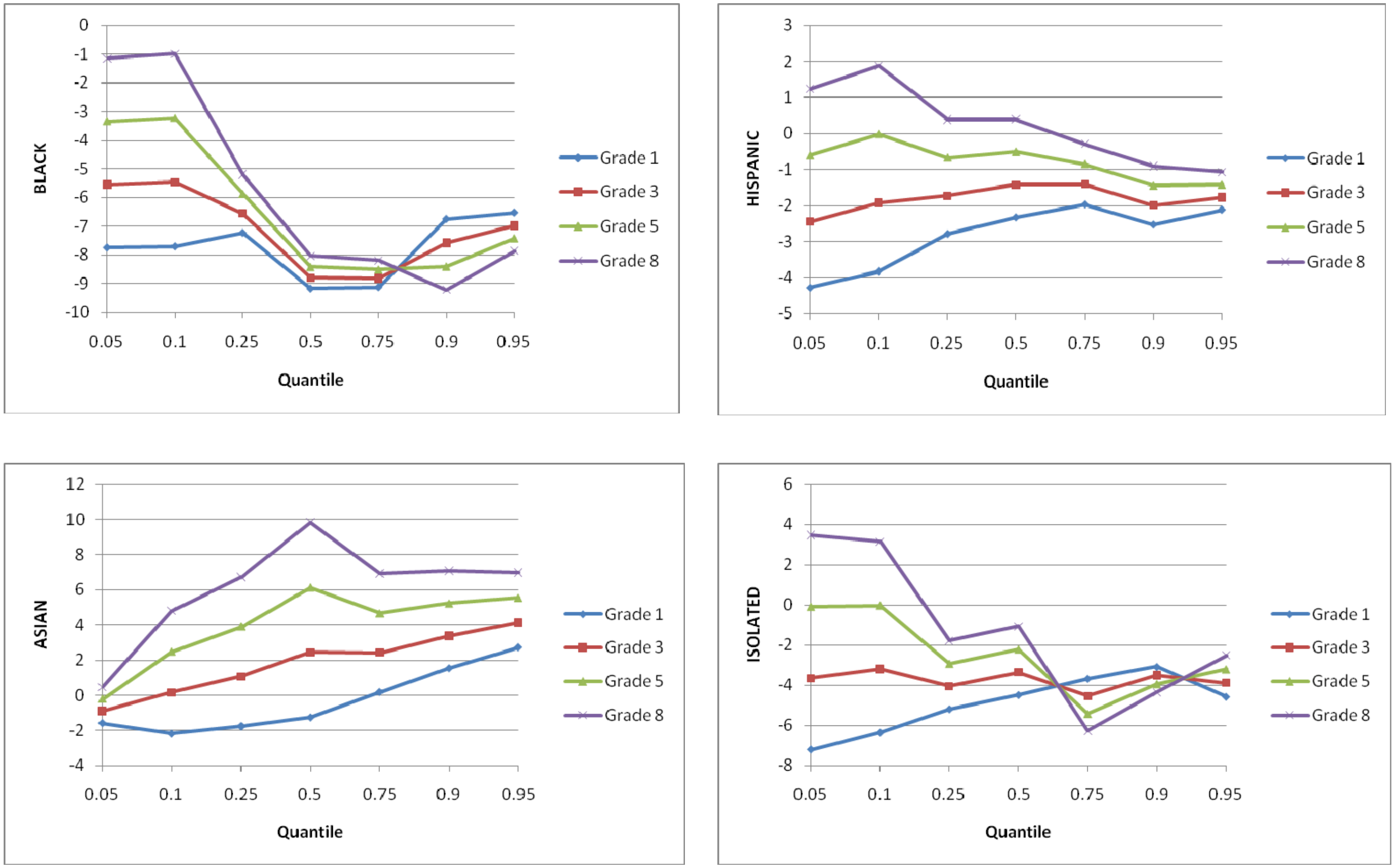


Figure 4.10 Race-Ethnicity Effect at Each Grade

Among all the race-ethnicity groups, Black students consistently scored lower than White students; and students at geographically isolated regions scored lower than their White counterparts at most ability levels. Interaction with TIME is significant at the lower tails of math score distribution for BLACK, HISPANIC and ISOLATED but at the central of the distribution for Asian. Figure 4.10 plots the race-ethnicity gaps across time for all four groups.

Although Black students performed significantly lower than White students at all grades, the gap is generally larger for high math ability students than for low math ability students. Thus, while the weaker ones are similarly weak between the two ethnicities, the stronger ones are more different between the ethnicities.

Figure 4.10 seems to suggest that Hispanic students are catching up with their White counterparts across time. By grade 8, lower math ability Hispanic students actually performed better than low math ability White students after language proficiency is controlled. Like Black students, however, at the high end of math score distribution, the Hispanic students still did not fare as well as the White students.

Asian is the only race-ethnicity group that not only catches up with the White by grade 3 at most quintiles, they actually outperformed the White students at grade 5 and 8 at all quantiles. The pattern of difference is similar to the Black and Hispanic versus the White: the gap between Asian and White is smaller for low ability students but larger for higher ability students. The only difference here is the direction of difference. In summary, the lower ability Asians caught up by grade 3 and the high ability Asians scored much higher than the high ability Whites at all grades.

Overall, the isolated students scored lower than the White students. Like Hispanics, at grade 8, the lower ability students in this group outperformed their White counterparts. The high ability students, however, still scored lower than their White peers. Like most other minority groups, there seems to be a ceiling that prevents the isolated group from being equal with the white at the high end of math score distribution.

READING*SES*TIME is significant at most quantiles and READING*SES is significant only at higher quantiles (Figure 4.7d). Figure 4.11plotted the READING effect at both the mean SES status and one standard deviation above the mean SES at each grade. The reference line corresponds to the former and the colored solid lines the latter. Clearly, for all grades, READING effect is the strongest at the central math ability distribution and weaker at both ends. This suggests if items are modified to minimize linguistic requirement in math assessment, a bigger score gain should be observed for the medium math ability students rather than the extremes. This pattern of impact is true with or without SES mediation in Figure 4.11. However, higher SES status seems to be able to downplay the language impact to some degree. And this moderating function of SES increases in magnitude from grade 1 to grade 8. This might be related to accumulated advantages of high SES families on math learning.

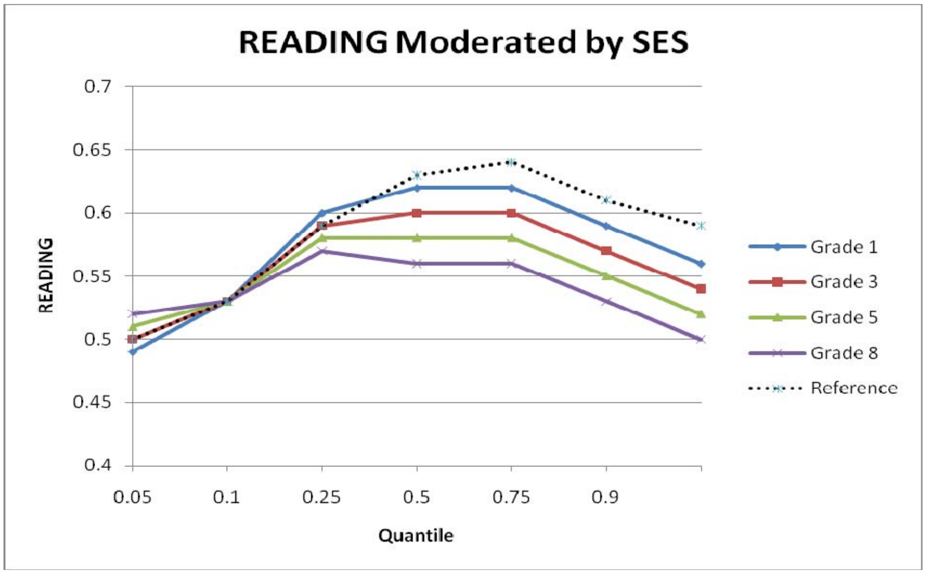


Figure 4.11 Differential READING Effect Moderated by SES at Each Grade

Note: The dotted line is the READING effect at the mean SES status. The colored solid lines refer to the calculated READING effect at each grade.

READING*GENDER*TIME is positive at all quantiles. READING*GENDER is negative and statistically significant at almost all quantiles. To better examine the READING*GENDER gap, Figure 4.12 plotted the differential READING impact for the two gender groups at each grade separately. Overall, the impact of language proficiency seems to be increasing as students' math ability increases. However, this trend stops at around the 75th percentile of conditional math distribution. From here on, the language impact started to decrease. Again, this could be that high math ability compensated for some deficiency of language skills.

More importantly, the READING impact varies between gender and grade. At grade 1, this impact is consistently stronger for boys than for girls. At later grades, this is only true for the lower math ability students (at or below 75th percentile at grade 3, at or below about the 40th percentile at grade 5 and at or below the 20th percentile at grade 8). At the other end of math ability, READING impact is stronger for girls than for boys. As Figure 4.7a has shown, girls consistently performed worse than boys in math, especially at the higher end of math score distribution. Their math ability may have not become strong enough to compensate for the language influence as it does for boys. Despite all these rippling effects, language has a consistent and positive relationship with math scores regardless of grade or gender.

Three-way and two-way interactions involving race-ethnicity are mostly significant. Figure 4.13 summarized the differential relationship between READING and MATH for each race-ethnicity group. Overall, READING has an increasing impact on math achievement for all groups until near the 90th percentile (for BLACK) or 75th percentile (for all other groups including WHITE). After this point, the impact starts to decrease. Furthermore, the impact increases as Black and Hispanic students grow, and decreases for the Asian students. There is no three-way interaction in the model involving the Hispanics based on previous decision during model building. Still, the impact is stronger for high math ability Hispanics than for low math ability Hispanics. These observations can be similarly explained by the compensatory and threshold hypotheses. Except Asians who might have passed some threshold and start to benefit from the compensation mechanism of their high math ability, the other minority groups all fell behind their White peers.

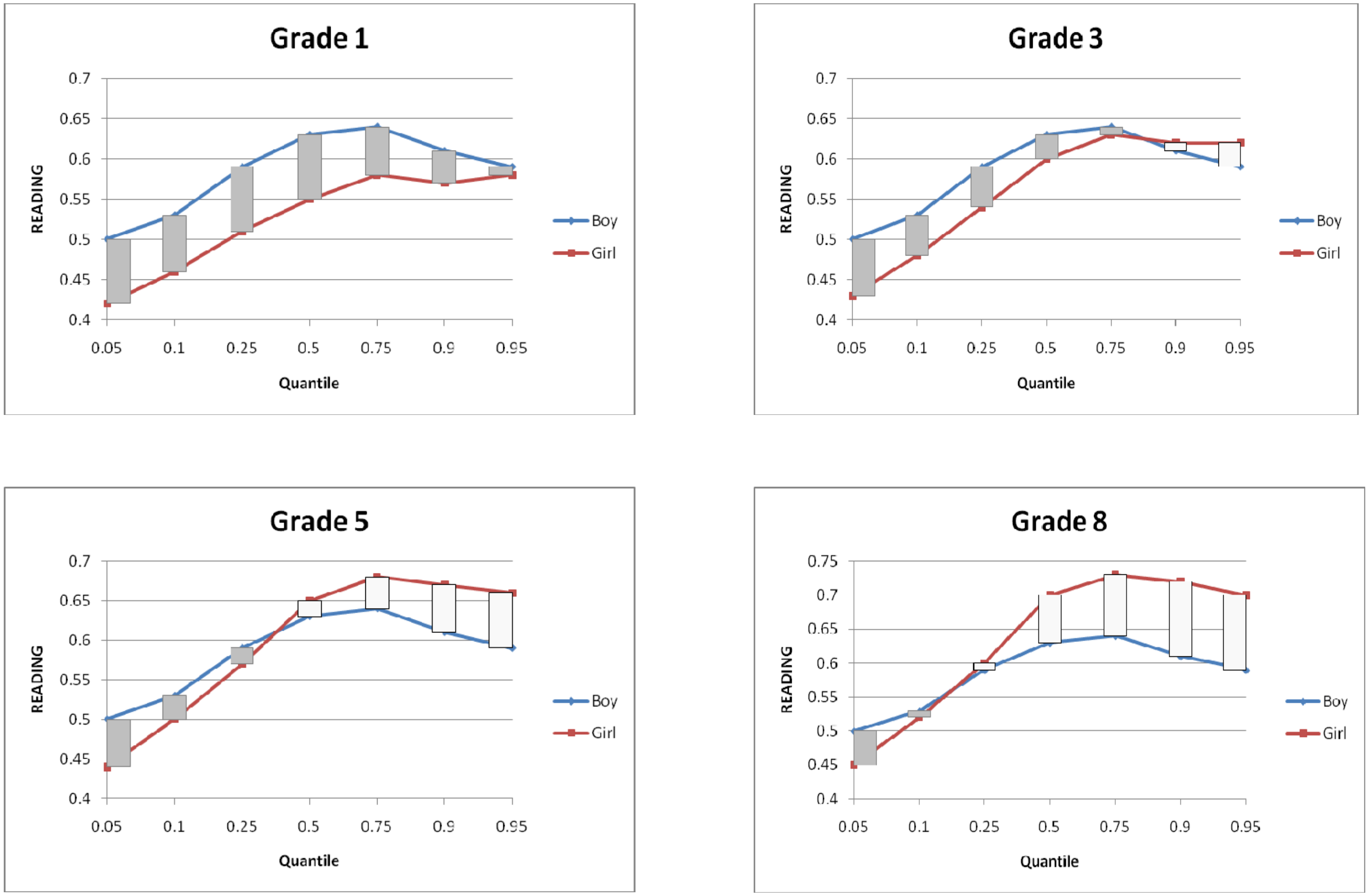


Figure 4.12 Differential READING Effect between Gender at Each Grade

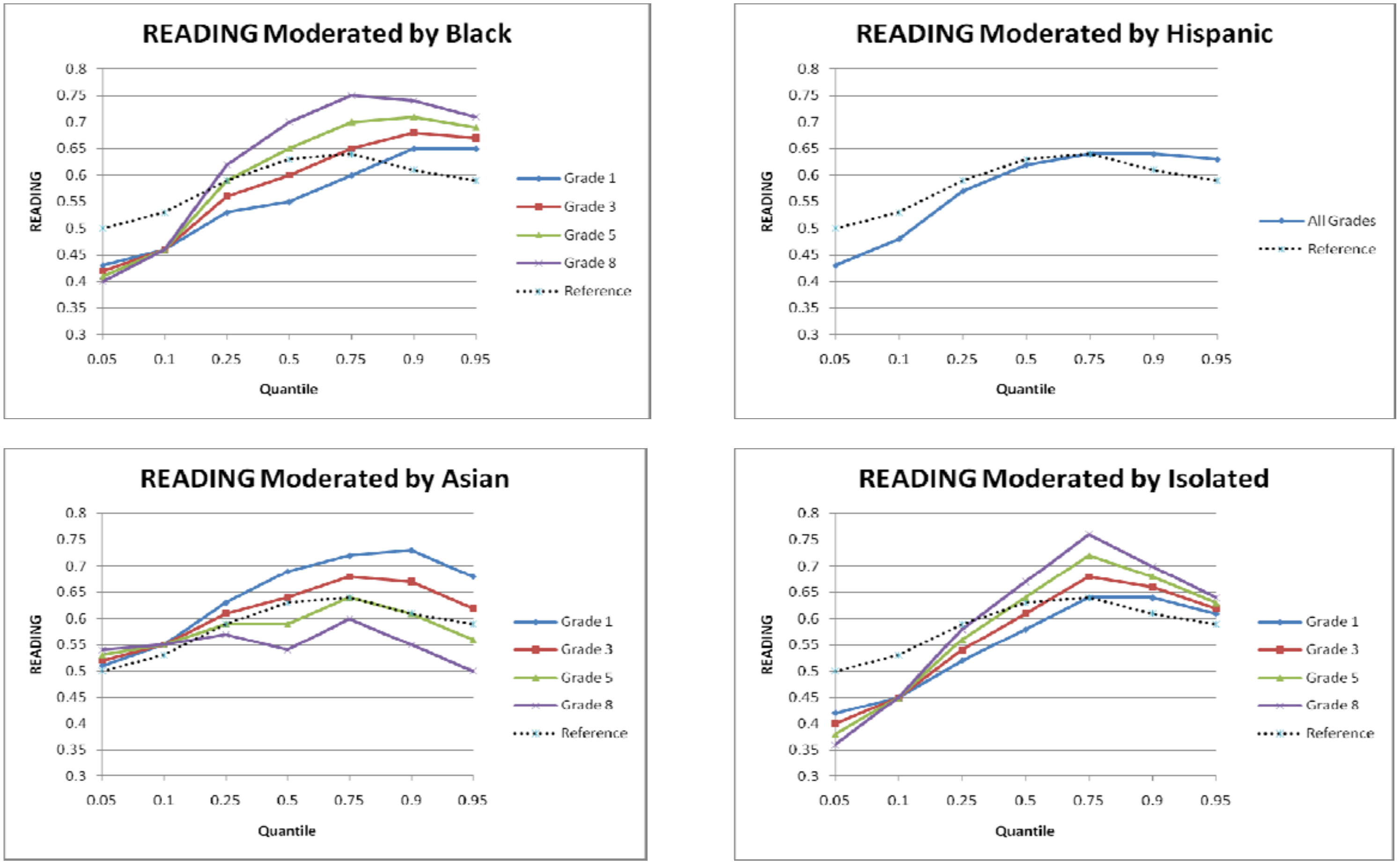


Figure 4.13 Differential READING Effect by Race at Each Grade

Summary

After individual level language proficiency is controlled, difference between ELLs, former ELLs and Non-ELLs disappear. Also, language proficiency has a consistent and positive effect on students' math performance. However, this effect is different along the math score distribution. Overall, the relationship is stronger for low math ability students and weaker for high math ability students. These answered research questions 1 and 3.

The answer to research question 2 is more complicated. In general, the results are consistent with past research on these variables. For example, there are obvious math achievement gaps between race-ethnicity groups. Except for Asian who caught up with the White reference group very quickly and outperformed the latter by grade 3, all other race groups are generally at or behind their White peers. Girls scored lower than boys at all quantiles within all grades and low SES students consistently performed worse than high SES students.

As expected, quantile regression provided richer information than mean regression, such as the differential language impact on math achievement depending on the math ability of the students. It also detected various, rather than constant, math achievement gaps related to gender, SES, or race-ethnicity. It is with the help of quantile regression that evidence is found that while gaps between low ability students decreased along time, there seem to be a ceiling for the disadvantaged (girls and/or most minority ethnicity groups) that prevented them from being comparable to the reference group at higher math ability. These are all the advantages that quantile regressions bring to the current research that mean regression cannot capture accurately.

It is prevalent through all the results that high math ability students seem to be less influenced by factors such as gender, SES or language proficiency. There are three hypotheses that can be adjusted to explain phenomena like the ones observed in the current study. Cummins (1979b) proposed a "threshold of linguistic competence" for bilingual students. He hypothesized that bilingual students may have to master a certain level of language proficiency before language is no longer a barrier for cognitive growth. Burns et al (1981) proposed a "technical threshold" in which they hypothesized students have to master technical language or symbolic language in the technical domain (here, math) before becoming good problem solvers. In the current data, the positive relationship between language and math means that high math ability students also tend to be high language ability students. According to the two threshold hypotheses, the decreasing language influence on math may be due to the fact that these students have passed the language threshold. More specifically, they have passed the technical threshold which is embedded in the definition of academic language for math. These threshold hypotheses can provide some degree of explanation for many rippling effects observed between language and math.

A third hypothesis, that is, *compensatory hypothesis* can also explain the discoveries regarding the moderating effect of most demographic variables on the relationship between language and math. The concept relevant to this hypothesis was first proposed by Meyer & Schvaneveldt (1971), later appeared in Posner and Snyder (1975a, 1975b) and then elaborated by Stanovich (1980). This hypothesis was originally used to explain the paradoxical patterns of fluent reading which neither a top-down or bottom-up model can explain cleanly. The terminology is adopted here to refer to a possible compensatory mechanism that may explain the observations that high math ability students (mostly at or above the 75th percentile) seem to be less sensitive to language impact within each gender or race group and with similar SES background. In other words, high math ability may be able to compensate for the disadvantage of language limitation regardless of other factors such as gender, SES and race-ethnicity. Why and what makes high ability students less sensitive to language influence may differ between the factors of gender, SES and race-ethnicity. In many cases, it might be the combination of both threshold and compensatory hypothesis that have been working jointly to offset these individual level disadvantages.

Chapter V

DISCUSSION

Implications

When controlling for language ability, the achievement gap between ELLs and Non-ELLs disappeared. This might reduce some stress for math teachers and school administrators. Also, it can be suggested that before a valid accountability decision is discussed, reporting academic achievement adjusted in terms of language proficiency may provide more accurate information on ELLs.

Results have shown that even between the two main ELL population source groups (Asian and Hispanic), language impact on math achievement is different. This implies that trying to accommodate the needs of various ELLs with a one-size-fits-all method for all groups of ELLs is inappropriate. Literature has already shown that only linguistic accommodations make a difference in ELLs' performance in assessments (Abedi & Hejri, 2004; Francis et al., 2006). This research exposed additional challenges in test accommodation. Just like LaCelle-Peterson and Rivera (1994) described: all ELLs vary in many other characteristics including cultural heritage, ethnic group affiliation, gender and individual learning differences. All these are educationally relevant and should be considered for instruction and assessment accommodations.

Many decisions are based on test scores. However, differences in scores may reflect other factors beyond students' competency (LaCelle-Peterson & Rivera, 1994; Minicucci & Olsen, 1992; Stevens, 1993). For example, if students did not have the opportunity to learn, that is, no access to the content knowledge and did not participate in classroom instruction and interaction, assessments based on equal opportunity to learn is biased. No statistical technique can correct for this type of inadequacy in assessment.

There is also implication for teacher preparation. Although the focus of the current study is assessment, lack of instruction preparation also harms students with low language proficiency during the learning process. The observed changing relationship between language and math may reflect the teachers' varying instructional styles in addition to the influence of language demand in assessment. The report by National Council on Teacher Quality (NCTQ, 2009) summarized that few states have prepared their elementary teachers well enough to teach reading and math. Only five states have an adequate test in reading instruction and only one state has an adequate test of mathematics. For middle schools, 21 states permit middle school teachers to teach on a k-8 generalist license. NCTQ concluded that this suggested many states believe that the pedagogy needed to teach later grade content is the same as the early elementary grades. The shifting relationship between READING and MATH in this study questions that practice. If teachers are not aware of the strong and shifting relationship between reading and math and have not been trained systematically to teach the two, their instruction may not enable differential approaches to help the linguistically disadvantaged. Stone (1988) speculated on the teaching of the deaf and made a statement that can be borrowed verbatim for the teaching of ELLs:

> It is possible that in the face of difficult communication, we have a tendency to eliminate as much conventional language as possible… Elimination of written and spoken language hardly enriched the context of instruction for students who have language problems, although in the short run this may seem to ease the task of the teacher. It is human nature to try to make a complex situation more simple, but finally the simple reduction of language in situations where communication is difficult … does not result in greater learning. (p.120)

Integrated teaching is a common practice in many classrooms now. However, traditional integrated teaching plans focus on a common topic spanning different disciplines at the same time. Results from the current research suggests to plan integrated teaching by paying attention to the academic language suitable for the topic but specific to each disciplines to best promote the learning. Rather than organizing units under weekly or monthly topics, using the academic language for the same topic but from different discipline will better help students to learn the content as well as the language. This is closely related to researches on academic English which is still at its infant stage.

Observations on grade 1 are very different from all other grades. For example, the relationship between language and math is increasing until the 75th percentile of conditional math ability at grade 1 but decreasing at all other grades; the variance of math score distribution is larger at high language proficiency and smaller at low language proficiency at grade 1 but the opposite for later grades; the relationship between SES and math keeps on increasing at grade 1 but is quite constant at all later grades; the confidence band for the slope estimate of ELL is too big at grade 1 but much smaller across other grades; and the pattern of significance change of race-ethnicity gap with or without READING controlled for at grade 1 is different from later grades. Performance at grade 1 is very different. The unique patterns for grade 1 seem to suggest that the construct for grade 1 may be more difficult to define thus leading to not so-well-guided test item development. Testing at grade 1 may not be as meaningful and vertical scaling may work better by excluding first grade.

The READING slope estimates started to decrease at the 75th percentile of conditional math ability at grade 1 and grade 3 for both boys and girls, students with various SES background and most race-ethnicity groups. It seems that the 75th percentile is a threshold. It is also starting from here, students' math ability seems to be able to compensate for the language influence in math assessment. What makes this specific percentile so important is not clear but future research may help locate the reasons and help guide instruction and assessment.

Limitations and Future Research

This is an exploratory research that needs to be cross-validated using different samples and instruments. Also, there are some limitations that future research should try to overcome.

Due to the limited number of ELLs and Former ELLs from different race-ethnicity groups, interaction terms between ELL status and race-ethnicity group were not studied. Future research may include these interaction terms to better control for the differences between various race-ethnicity groups. Interaction between other covariates such as SES and GENDER or SES and ELL status are not of interest to the key issue here. However, future researches may also include these interaction terms to explore other issues. The requirement of sample size is a major consideration.

One justification for the choice of quantile regression over mean regression in the current study is the nonnormal error distribution. However, non-normality can also be due to model misspecification (Fox, 1997). R^2 and pseudo-R^2 show that the covariates explained only half of the variance in the math scores. These means that there are other variables that could be included if the purpose is to explain the math score variance. Literature has suggested other factors such as participation in advanced courses (Myers & Milne, 1988) and the number of courses taken (Mau & Lynn, 2000; Tate, 1997) in math are important factors for that purpose. Once included, these may interact with the relationship between language proficiency and math achievement.

It should be pointed out that quantile regression is not the only robust model when normal assumption or homoscedasticity does not hold. When normality holds, homoscedasticity can be modeled by including a shifting error term. However, this function is not widely implemented in commonly available packages. As demonstrated, quantile regression method is straightforward and graphic presentation of results can facilitate non-statisticians to use the model rather than limit it to "experts." Quantile regression modeling can provide much more information beyond a regular mean regression. The interpretation, however, is very challenging. This might be a natural dilemma for any models: the more complicated it is, the more information that can be gleaned but the more difficult to summarize.

As Monroe and Englehart (1931) pointed out long ago, to better inform instruction and assessment, it is urgent to study the relationship between specific language skills and math skills. If the specific language skills that affect math achievement at each grade can be identified, this may guide math instruction and test development tremendously. Grimm (2008) explored the relationship between reading comprehension and three components of mathematics. He found out that reading has a strong relationship with Problem Solving and Data Interpretation but is almost irrelevant to Mathematical Computation. His research used a general reading test (Iowa Test of Basic Skills), however, it is reasonable to hypothesize that not all reading skills are equally relevant to the above math components during learning or assessment. For example, the complexity of sentence, specialized vocabulary, genre and language function for math may vary across grades (Bailey & Butler, 2003); thus a general reading skill measurement has a different effect than academic language proficiency on math achievement. Actually, both math sub-skills and reading sub-skills are identified in the ECLS-K data for the current study, but the high multicollinearity within each subject prevented the exploration of detailed relationship in this direction. This is an important issue that deserves more research in the future.

English language proficiency in the current study is treated as an exogenous variable that affects the math outcome. Literature exists that has also explored factors that contribute to the difference in second language proficiency such as native language proficiency (Oller, 1980), overlap in oral and written form between first language and second language (Dressler, 2006), language aptitude (Clément & Kruidenier, 1985), learner strategies (O'Malley & Chamot, 1990; Oxford, 1989), motivation (Gardner, 1988; Ramage, 1990) and learner type (Skehan, 1991). In that case, English language proficiency is also an endogenous variable influenced by other factors. This is the quantile regression version of structural equation modeling (SEM). Once developed, it can give even more insight to the relationship between these factors, language proficiency and math achievement. Methodological research in this direction has already started (Koenker, 2005). Application to language testing may be possible in the near future.

Related to the methodological research of SEM, it is also fundamental to understand the issue of multicollinearity in the context of quantile regression. Indices such as vector inflation factor (VIF) are based on mean regression and ordinary least squares. Because quantile regression minimizes weighted absolute deviation rather than squares of deviation, the concept of multicollinearity takes on a different meaning. How to diagnose collinearity in quantile regression may also be different. Whether mean regression modeling is more sensitive to multicollinearity than quantile regressions is an interesting and important line of research. Research on this will advance understanding and application of quantile regression.

A limitation not specific to this study but to longitudinal studies in general is that students are not representative of current grades other than the one where they were first identified. The grade level gaps between gender and race-ethnicity groups might not be the same as any current grade randomly chosen in the U.S. at this moment. Even the results based on another longitudinal study may be different from this study. There are too many factors involved in any research, of which sampling framework and time are just two examples. Thus, the specific estimates in this study should be used with caution.

REFERENCES

Abedi, J. (1999a). *NAEP math test accommodations for students with limited English proficiency*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. (1999b). *The impact of student's background characteristics on accommodation results for students with limited English proficiency*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*, 231-257.

Abedi, J. (2007). *English language proficiency assessment in the nation: Current status and future practice*. Davis, CA: University of California – Davis.

Abedi, J., & Hejri, F. (2004). Accommodations in the national assessment of educational progress for students with limited English proficiency. *Applied Measurement in Education, 17*, 371-392.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-134.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1-28.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. et al (2005). *Validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (Final Deliverable to OERI/OBEMLA; pp. 1-45). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abreveya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, *26*, 247-257.

AERA, APA, NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Aiken, L. R. (1971). Verbal factors and mathematics learning: A review of research. *Journal for Research in Mathematics Education, 2*(4), 304–313.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Education Research, 42,* 359–385.

Astin, A.W. (1982). *Minorities in American higher education.* San Francisco: Jossey-Bass.

Austin, P., Tu, J., Daly, P., & Alter, D. (2005). The use of quantile regression in health care research: A case study examining gender differences in the timeliness of thrombolytic therapy. *Statistics in Medicine, 24,* 791-816.

Ausubel, D. P., & Robinson, F. G. (1969). *School learning: An introduction to educational psychology*. New York: Holt, Rinehart & Winston.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, A. L. (2000). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation From Three Perspectives* (Final Deliverable to OERI/OBEMLA; pp. 85–115). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to k-12 education: A Design Document.*

CSE Report 611. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., Butler, F. A., & Sato, E. (2005). *Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards*. (CSE Report 667). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Balow, I. H. (1964). Reading and computation ability as determinants of problem solving. *Arithmetic Teacher, 11,* 18-22.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210,* 1262-1264.

Benbow, C. P. (1992). Academic achievement in mathematics and science of students between ages 13 and 23: Are there differences among students in the top one percent of mathematical ability? *Journal of Educational Psychology, 84,* 51–61.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and practice, 28*(4), 42-51.

Brown, C. L. (2005). Equity of literacy-based math performance assessments for English language leaners. *Bilingual Research Journal, 29*(2), 337-497.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica, 62*(2), 405-458.

Burns, M., Gerace, W., Mestre, J. P., & Robinson, H. (1983). The current status of Hispanic technical professionals: How can we improve recruitment and retention. *Integrateducation, 20,* 49-55.

Butler, F. A., & Castellon-Wellington, M. (2005). Students' concurrent performance on tests of English language proficiency and academic achievement. In *Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation from Three Perspectives* (Final Deliverable to OERI/OBEMLA; pp. 47-83). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment, 1*(8), 412-420.

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. Brennan (Ed.), *Educational measurement* (pp. 517-530). Westport, CT: American Council on Education and Praeger Publishers.

Chalhoub-Deville, M., & Deville, C. (2008). Nationally mandated testing for accountability: English language learners in the US. In B. Spolsky & F. Hult (Eds.), *Handbook of educational linguistics* (pp. 510–522). Oxford, England: Blackwell.

Chang, M., Singh, K., & Filer, K. (2009). Language factors associated with achievement grouping in math classrooms: a cross-sectional and longitudinal study. *School Effectiveness and School Improvement, 20*(1), 27- 45.

Chen, C. (2007). A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics, 16,* 136–164.

Chevapatrakul, T., Kim, T., & Mizen, P. (2009). The Taylor Principle and monetary policy approaching a zero bound on nominal rates: Quantile regression results for the United States and Japan. *Journal of Money, Credit and Banking, 41*(8), 1705-1723.

Clément, R., & Kruidenier, B. (1985). Aptitude, attitude, and motivation in second language proficiency: A test of Clement's model. *Journal of Language and Social Psychology, 4,* 21-37.

Cole, T.J. (1988). Fitting smoothed centile curves to reference data (with discussion). *Journal of the Royal Statistical Society, Series A (Statistics in Society), 151*(3), 385-418.

Collier, V. P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly, 21,* 617-641.

Cottrell, R. S.(1968). *A study of selected language factors associated with arithmetic achievement of third grade students.* (Doctoral dissertation, Syracuse University) Ann Arbor, Mich.: University Microfilms.

Cummins, J. (1979a). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19,*121-129.

Cummins, J. (1979b). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49,* 222-251.

Cummins, J. (1981a). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics, 2,* 132-149.

Cummins, J. (1981b). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework.* Evaluation, Dissemination and Assessment Center, California State University, Los Angeles.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, England: Multilingual Matters.

Cummins, J. (1999). *BICS and CALPS: Clarifying the distinction*. Retrieved on October 13th, 2009 from http://eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED438551&ERICExtSearch_SearchType_0=no&accno=ED438551

Davidson, F. (1994). Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms. *Language Testing, 11,* 83-95.

Dressler, C. (2006). First- and second-language literacy. In D. L. August & T. Shanahan (Eds.), *Developing literacy in a second language: Report of the National Literacy Panel*. Mahwah, NJ: Lawrence Erlbaum Associates.

Edgeworth, F. (1888). On a new method of reducing observations relating to the several quantiles. *Philosophical Magazine, 25,* 184-191.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1-26.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage Publications.

Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical Guidelines for the Education of English Language Learners: Research-Based Recommendations for the Use of Accommodations in Large-Scale Assessments.* (Under cooperative agreement grant S283B050034 for U.S. Department of Education). Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Freeman, B., & Crawford, L. (2008). Creating a middle school mathematics curriculum for English-language learners. *Remedial and Special Education, 29*(1), 9-19.

Fry, R. (2007). *How far behind in math and reading are English language learners?* Washington, DC: Pew Research Center. Retrieved on May 28, 2010, from http://pewhispanic.org/files/reports/76.pdf

Galindo, C. (2009). *English language learners' math and reading achievement trajectories in the elementary grades: Full technical report*. Retrieved online on May 28th, 2009 from http://nieer.org/resources/research/English_language_learners_math_reading_achievement_trajectories_elem.pdf.

Galindo, C. (2010). English language learners' math and reading achievement trajectories in the elementary grades. In E. García & E. Frede (Eds.), *Developing the research agenda for young English language learners* (pp. 42-58). NYC: Teachers College Press.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Calahan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75,*165– 190.

Garcia, G.N. (2000). *Lessons from research: What is the length of time it takes limited English proficient students to acquire English and succeed in an all-English classroom?* National Clearinghouse for Bilingual Education Issue Brief (Washington, DC) VOL (5.)

Gardner, R.C. (1988). The socio-educational model of second language learning: Assumptions, findings, and issues. *Language Learning, 38,* 101-126.

Geary, D. C. (1994). *Children's mathematical development*. Washington DC: American Psychological Association.

Ginsburg, H. P., & Russell, R. L. (1981). Social class and racial influences on early mathematical thinking. *Monographs of the Society for Research in Child Development*, 46.

Goh, D. S. (2004). *Assessment accommodations for diverse learners*. Boston: Pearson Education.

Good, P.I., & Hardin, J.W. (2006). *Common errors in statistics (and how to avoid them)*, 2nd Ed. Hoboken, New Jersey: John Wiley & Sons.

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology, 33*(3), 410-426.

Haile G.A., & Nguyen, A.N. (2008). Determinants of academic attainment in the United States: A quantile regression analysis of test scores. *Educational Economics, 16* (1), 29-57.

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stake testing. *Educational Measurement: Issues and Practice, 23* (1), 17-27.

Hao, L., & Naiman, D.Q.(2007). *Quantile regression*. Thousand Oaks, CA: SAGE publications.

Han, W. J. (2008). The academic trajectories of children of immigrants and their school environments. *Developmental Psychology, 44*(6), 1572-1590.

He, X., & Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association, 97*, 783–795.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161.

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

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155.

Kaplan, B. J., & Weisberg, F. B. (1987). Sex differences and practice effects on two visual–spatial tasks. *Perceptual and Motor Skills, 64*, 139–142.

Kato, K., Albus, D., Liu, K., Guven, K., & Thurlow, M., (2004). *Relationships between a statewide language proficiency test and academic achievement assessments*. (LEP Projects Report 4). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.

Khmaladez, E.V.(1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability and its Applications, 26*, 240-257.

Kieffer, M.J., Lesaux, N.K., Rivera, M., & Francis, D.J. (2009). Effectiveness of accommodations for English Language Learners taking large-scale assessments. *Review of Education Research, 79*(3), 1168-1201.

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics, 14*, 41-55.

Koenker, R. (2005). Quantile regression, Econometric Society Monographs (48). *Econometric Society,* Cambridge University Press.

Koenker, R. (2009). Quantreg: Quantile regression. R package version 4.27. Retrieved on May 5th, 2010 from http://CRAN.R-project.org/package=quantreg

Koenker, R. & Bilias, Y. (2001). Quantile regression for duration data: A reappraisal of the Pennsylvania Reemployment Bonus Experiments. *Empirical Economics, 26*, 199-220.

Koenker, R., & d'Orey, V. (1994). Remark AS R92: A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores. *Applied Statistics, 43*, 410–414.

Koenker R. & Machado J.A.F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association, 94*(448), 1296-1310.

Koenker, R., & Geling, R. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of American Statistical Association, 96*, 458-468.

Koenker, R., & Hallock, K.F. (2001). Quantile regression. *The Journal of Economic Perspectives, 15*(4), 143-156.

Koenker, R., & Xiao, Z. (2002). Inference on the quantile regression process. E*conometrica, 81*, 1583–1612.

Kopriva, R. J., Bauman, J., Cameron, C., & Triscari, R. (2009). *Final research report: Obtaining necessary parity through academic rigor in science*. PR/Award # 368A06007. Center for Applied Linguistics, Washington DC.

Kretschmer, R. E. (1991). Exceptionality and the limited English proficient student: Historical and practical contexts. In E. V. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students* (pp.1-38). Austin, TX: Pro-Ed.

LaCelle-Peterson, M.W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*, 55-75.

Lam, T.C.M. (1993). Testability: A critical issued in testing language minority students with the standardized achievement tests. *Measurements and Evaluation in Counseling and Development, 26*, 179–191.

Leahey, E., & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces, 80*, 713– 732.

Levin, J. (2001). For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. *Empirical Economics, 26*, 221-246.

Liu, O. L. & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*, 164-184.

Lummis, M., & Stevenson, H. W. (1990). Gender differences in beliefs and achievement: A cross-cultural study. *Developmental Psychology, 26*, 254–263.

Martínez, J. F., Bailey, A. L., Kerr, D., Huang, B. H., & Beauregard, S. (2009). *Measuring opportunity to learn and academic language exposure for English language learners in elementary science classrooms.* (CRESST Report 767). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mau, W. C., & Lynn, R. (2000). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Psychology, Evolution and Gender, 2*, 119– 125.

Menken, K. (2000). *What are the critical issues in wide-scale assessment of English language learners?* NCSB Issues & Briefs, 6, 1-4. Washington, DC: National Clearinghouse for Bilingual Education.

Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200–220). Hillsdale, NJ: Erlbaum.

Minicucci, C., & Olsen, L. (1992). *Programs for secondary limited English proficient students: A California study* (Occasional Papers in Bilingual Education, No. 5). Washington, DC: National Clearinghouse for Bilingual Education.

Monroe, W. S., & Englehart, M. D. (1931). A critical summary of research relating to the teaching of arithmetic. *Bureau of Educational Research Bulletin, 58*, University of Illinois.

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227-234.

Myers, D. E. (1988). Effects of home language and primary language on mathematics achievement: A model and results for secondary analysis. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 259-293). Hillsdale, NJ: Erlbaum.

Najarian, M., Pollack, J. M., & Sorongon, A.G. (2009). *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K), psychometric report for the eighth grade (NCES 2009–002).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

National Council on Teacher Quality. (2009). *2009 state teacher policy yearbook*, Retrieved on September 13th, 2010 from http://www.nctq.org/stpy09/reports/stpy_national.pdf

Oller, J. W. Jr. (1980). A language factor deeper than speech: More data and theory for bilingual assessment. In J. E. Alatis (Ed.), *Current issues in bilingual education*. Georgetown University Roundtable in Languages and linguistics (pp.14-30). Washington, DC: Georgetown University Press.

O'Malley, J. M., & Chamot, A.U. (1990). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.

Oxford, R. (1989). *Language learning strategies: What every teacher should know*. Rowley, MA: Newbury House.

Pandey, G. R., & Nguyen, V.T. (1999). A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology, 225*, 92-101.

Parzen, M. I., Wei, L.J. & Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika, 81*, 341–350.

Pennock-Román, M. (1990). *Test validity and language background*. New York: College Entrance Examination Board.

Portnoy, S., & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computation of squared-error vs. absolute-error estimators. *Statistical Science, 12*, 279–300.

Posner, M.I., & Snyder, C.R.R. (1975a). Attention and cognition control. In R. Solso (Ed.), *Information processing and cognition: The Loyola symposium*. Hillsdale, N.J.: Erlbaum Associates.

Posner, M. I., & Snyder, C. R. R. (1975b). Facilitation and inhibition in the processing of signals. In P. M. A Rabbitt & S. Dornic (Eds.), *Attention and performance V*. New York: Academic Press.

Powell, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. In W. Barnett, J. Powell, & G. Tauchen (Ed). *Nonparametric and semparamtric methods in econometrics*. Cambridge: Cambridge University press.

Prieto-Rodriguez, J., Barros, C. P., & Vieira, J. A. (2008). What a quantile approach can tell us about returns to education in Europe. *Education Economics, 16*(4), 391-410.

Ramage, K. (1990). Motivational factors and persistence in foreign language study. *Language Learning, 40*, 189-219.

Rivera, C., Collum, E., & Shafer Willner, L. (Eds.). (2006). *State assessment policy and practice for English language learners: A national perspective.* Mahwah, NJ: Lawrence Erlbaum.

SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide.* Cary, NC: SAS Institute Inc.

Secada,W. G. (1992). Race, ethnicity, social class, language, and achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 623–660). New York: Macmillan.

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition, 13*, 275-298.

Stanovich, K.E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*(1), 32-71.

Stevens, F.I. (1993). *Opportunity to learn: Issues of equity for poor and minority students*. Washington, DC: National Center for Education Statistics.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of English language learners*. Los Angeles, CA: Universityof California, National Center for Research on Evaluation, Standards, and Student Testing.

Stone, J. B. (1988). Intention and convention in mathematics instruction: Reflections on the learning of deaf students. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 63-72). Hillsdale, NJ: Erlbaum.

Tate, W. F. (1997). Race–ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*(6), 652– 679.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K), combined user's manual for the ECLS-K eighth-grade and K–8 full sample data files and electronic codebooks* (NCES 2009–004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Tran, Z. (2005). *Help with English Language Proficiency "HELP"program evaluation of sheltered instruction multimedia lessons*. Retrieved on March 10th, 2010, from http://www.helpprogram.net

U. S. Department of Commerce, Bureau of Census (1993). *Statistical abstract of the United States: The national data book (115th ed)*. Washington, DC: U.S. Government Printing Office.

U.S. Department of Education, National Center for Education Statistics. (2008). *Mathematics achievement of language-minority students during the elementary years* (NCES 2009-036). Washington, DC: U.S Government Printing Office.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine, 25*, 1369-1382.

Wilkinson, L., & Silliman, E. (2000). Classroom language and literacy learning. In M. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research*, Vol. III (pp. 337-360). Mahwah, NJ: Lawrence Erlbaum Associates.

Wößmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics, 13*(2), 143-169.

Wright, W.E., & Li, X. (2008). High-stake math tests: How No Child Left Behind leaves newcomer English language learners behind. *Language Policy, 7*, 237-266.

Yang, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *Journal of American Statistical Association,, 94*, 137-145.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 273– 283). New York: Springer.

Yu, K., Lu, Z., & Stander, J. (2003).Quantile regression: Applications and current research areas. Journal of the Royal Statistical Society, Series D (The Statistician), 52(3), 331-350.

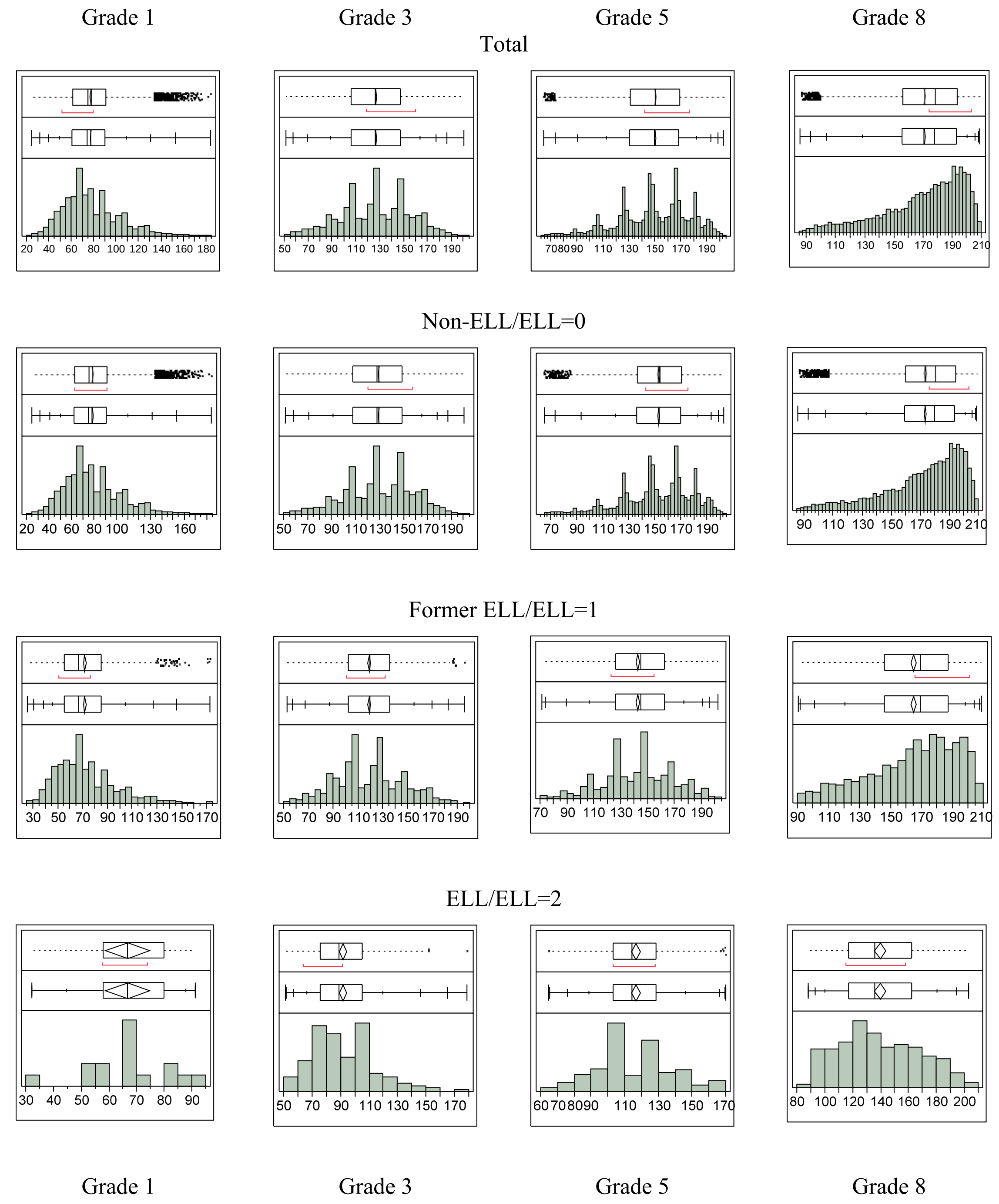
APPENDIX A: VARIABLE SUMMARY

Variable Name	Meaning	Scale	Possible values	Note
MATH	Math Scaled Score	Continuous	0 - 212	Students' responses are scaled by using all the responses for all 7 rounds. The final scores are thus on the same scale.
READING	Reading Scaled Score	Continuous	1 - 174	Students' responses are scaled by using all the responses for all 7 rounds. The final scores are thus on the same scale. For interpretation purpose, Reading scores are group centered for grade level analysis and grand-mean centered for longitudinal model.

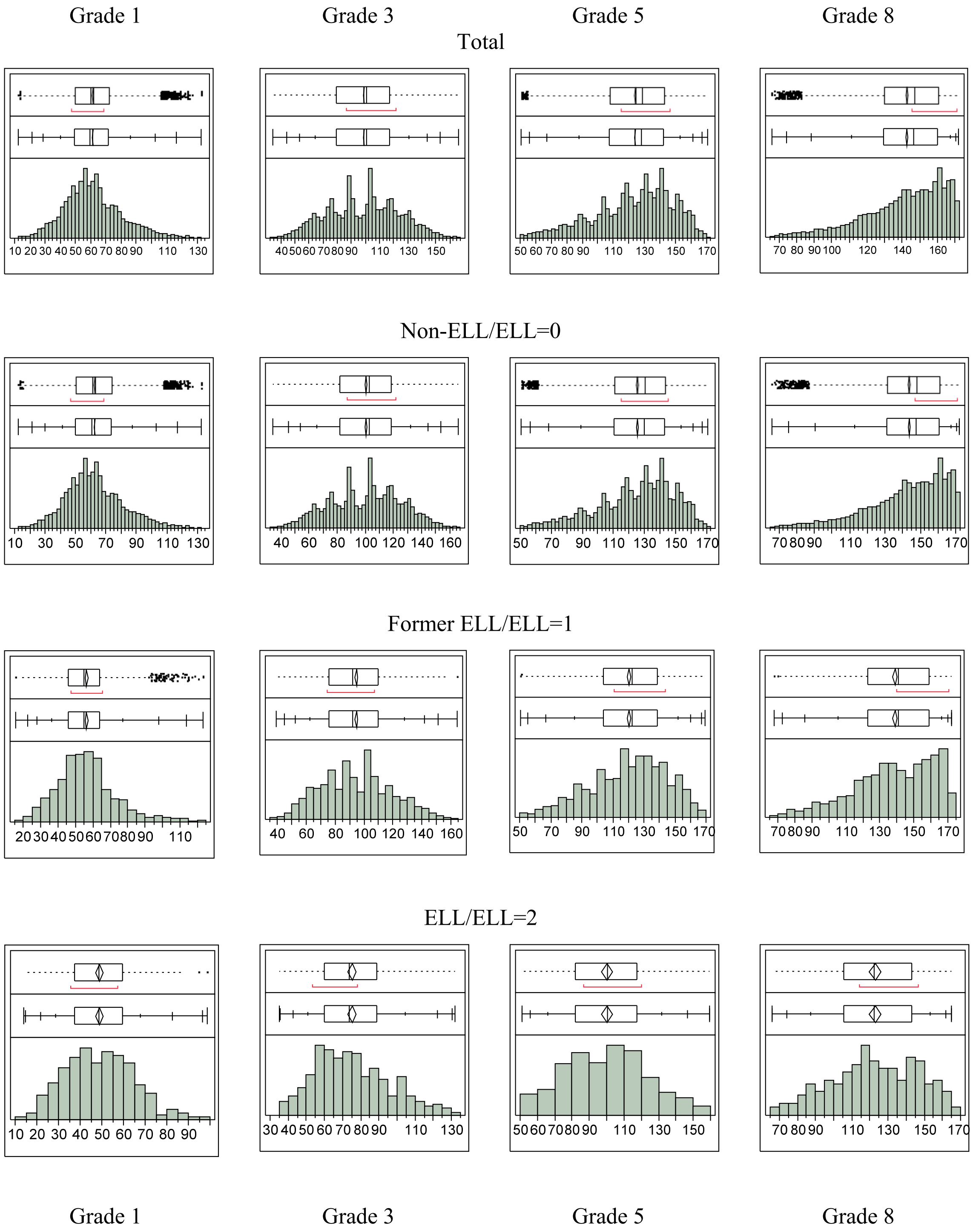
ELL	ELL status	Norminal	0, 1, 2	Non-ELL=0
				FormerELL=1
				ELL=2

GENDER	Gender	Norminal	0, 1	Male=0
				Female=1

SES	Family socioeconomic background	Continous	-2.48 to 2.54	A composite of parent's income, educational levels and occupations.
RACE	Race-ethnicity	Norminal	1 - 5	White=5 Black=4 Hispanic=3 Asian=2 Isolated=1 Isolated includes all native Americans such as Indians, Hawaiians, Pacific Islanders and Alaska natives. All the variables are also dummy coded as indicators for comparison purpose.



Reading score distribution



Math score distribution

APPENDIX B: DESCRIPTIVE AND DIAGNOSTIC STATISTICS

Descriptive Statistics

	Variable	Mean	Std. Deviation	N
G1	MATH	63.91	18.24	8072
	READING	2.52	23.83	8072
G3	MATH	102.05	24.22	8023
	READING	3.69	27.73	8023
G5	MATH	126.18	24.26	7992
	READING	2.82	26.02	7992
G8	MATH	143.41	21.56	7959
	READING	1.22	27.18	7959

Note: READING is centered around group mean.

		Correlations										VIF
		MATH	GENDER	FORMER	ELL	BLACK	HISPANIC	ASIAN	ISOLATED	SES	READING	
Grade 1	MATH	1.00										
	GENDER	-.06	1.00									1.01
	FORMER	-.15	.01	1.00								1.69
	ELL	-.01	.00	-.01	1.00							1.00
	BLACK	-.20	.01	-.11	.02	1.00						1.12
	HISPANIC	-.15	-.01	.50	.02	-.14	1.00					1.58
	ASIAN	.01	.02	.32	.03	-.08	-.10	1.00				1.27
	ISOLATED	-.10	.02	-.03	-.01	-.06	-.08	-.04	1.00			1.04
	SES	.40	-.01	-.21	-.01	-.20	-.23	.03	-.10	1.00		1.30
	READING	.65	.09	-.13	-.01	-.14	-.15	.08	-.09	.39	1.00	1.21

		Correlations										VIF
		MATH	GENDER	FORMER	ELL	BLACK	HISPANIC	ASIAN	ISOLATED	SES	READING	
Grade 3	MATH	1.00										
	GENDER	-.10	1.00									1.01
	FORMER	-.10	.00	1.00								1.69
	ELL	-.16	-.01	-.05	1.00							1.23
	BLACK	-.23	.01	-.10	-.04	1.00						1.14
	HISPANIC	-.18	-.01	.46	.30	-.15	1.00					1.74
	ASIAN	.04	.01	.32	.01	-.08	-.11	1.00				1.27
	ISOLATED	-.10	.02	-.02	-.03	-.06	-.08	-.04	1.00			1.05
	SES	.44	-.01	-.20	-.19	-.18	-.27	.03	-.09	1.00		1.40
	READING	.73	.08	-.14	-.22	-.19	-.22	.02	-.11	.47	1.00	1.39

		Correlations										VIF
		MATH	GENDER	FORMER	ELL	BLACK	HISPANIC	ASIAN	ISOLATED	SES	READING	
Grade 5	MATH	1.00										
	GENDER	-.11	1.00									1.01
	FORMER	-.08	.00	1.00								1.70
	ELL	-.15	-.01	-.05	1.00							1.24
	BLACK	-.25	.01	-.11	-.04	1.00						1.15
	HISPANIC	-.16	-.01	.46	.30	-.15	1.00					1.74
	ASIAN	.07	.02	.32	.01	-.08	-.11	1.00				1.28
	ISOLATED	-.10	.02	-.03	-.03	-.06	-.08	-.04	1.00			1.05
	SES	.45	-.01	-.20	-.20	-.19	-.27	.02	-.09	1.00		1.42
	READING	.73	.06	-.13	-.21	-.21	-.20	.01	-.12	.48	1.00	1.39

		Correlations										VIF
		MATH	GENDER	FORMER	ELL	BLACK	HISPANIC	ASIAN	ISOLATED	SES	READING	
Grade 8	MATH	1.00										
	GENDER	-.06	1.00									1.01
	FORMER	-.08	.01	1.00								1.67
	ELL	-.15	-.01	-.05	1.00							1.22
	BLACK	-.25	.01	-.11	-.04	1.00						1.16
	HISPANIC	-.16	-.01	.46	.30	-.15	1.00					1.74
	ASIAN	.08	.02	.31	.00	-.08	-.11	1.00				1.25
	ISOLATED	-.09	.02	-.03	-.03	-.06	-.08	-.04	1.00			1.04
	SES	.44	-.01	-.19	-.20	-.19	-.27	.03	-.08	1.00		1.41
	READING	.73	.09	-.10	-.19	-.25	-.20	.06	-.09	.47	1.00	1.40

APPENDIX C: GRADE LEVEL MODEL COMPARISON STATISTICS

Wald Test

		.05	.1	.25	.5	.75	.9	.95
Grade 1	S vs.F	140.74	206.17	384.07	618.08	848.77	471.41	379.88
	F vs.I	1570.57	2016.7	2378.99	2456.55	1284.91	1086.64	688.53
Grade 3	S vs.F	60	179.7	450.08	704.87	603.95	377.94	377.37
	F vs.I	973.28	1982.01	2868.9	3514.16	3585.54	1724.03	1352.42
Grade 5	S vs.F	217.25	222.74	608.65	560.52	636.1	371.97	314.32
	F vs.I	1609.88	2219.09	3716.25	3412.36	2644.35	1554.87	1179.39
Grade 8	S vs.F	102.5	223.58	511.91	383.45	377.03	215.4	92.62
	F vs.I	1455.62	2089.08	3051.81	3054.65	3260.81	1737.36	771.14

Likelihood Ratio Test

		.05	.1	.25	.5	.75	.9	.95
Grade 1	S vs.F	134.26	190.53	295.11	479.06	552.02	411.68	263.44
	F vs.I	954.75	1177.88	1754.46	2100.88	1953.35	1231.53	767.42
Grade 3	S vs.F	81.92	158.7	355.31	639.82	582.31	352.11	270.44
	F vs.I	837.66	1338.38	2064.98	2707.49	2317.39	1479.2	1144.07
Grade 5	S vs.F	166.64	201.83	420.98	623.04	618.58	372.56	256.71
	F vs.I	990.19	1243.11	2100.39	2446.4	2229.27	1372.89	874.88
Grade 8	S vs.F	105.19	204.84	411.49	389.69	341	190.91	127.79
	F vs.I	966.21	1390.59	2065.59	2428.86	2231.57	1478.91	1011.16

Note:
S refers to simple regression where there is only one independent variable: Reading.
F refers to a full regression with all other covariates in addition to Reading but without any interactions.
I refers to the full regression with all the interaction terms in addition to main effects.
All the statistics are statistically significant below .01 level.

APPENDIX D: GRADE LEVEL FULL MODEL

Table 4.2a. Grade 1

	QRM							MRM
	.05	.1	.25	.5	.75	.9	.95	
Intercept	45	49.33	56.27	64.9	75.1	85.16	92.53	66.4
	(.38)	(.32)	(.27)	(.31)	(.41)	(.49)	(.69)	(.24)
	***	***	***	***	***	***	***	***
READING	.41	.43	.48	.51	.56	.59	.58	.5
	(.01)	(.01)	(.01)	(.01)	(.02)	(.03)	(.04)	(.01)
	***	***	***	***	***	***	***	***
SES	2.18	2.23	2.4	2.9	3.79	3.65	3.96	3.15
	(.32)	(.29)	(.22)	(.23)	(.32)	(.48)	(.64)	(.21)
	***	***	***	***	***	***	***	***
GENDER	-.56	-1.3	-2.27	-3.61	-4.98	-6.12	-7.35	-3.76
	(.45)	(.39)	(.31)	(.32)	(.45)	(.6)	(.8)	(.3)
		***	***	***	***	***	***	***
FORMER	-.87	-.98	-.14	-.11	-.92	-.72	-1.46	-.72
	(1.06)	(.79)	(.63)	(.6)	(.87)	(1.57)	(2.14)	(.59)
ELL	12.7	9.23	2.98	.24	-7.64	-15.63	-22.25	-2.24
	(359.26)	(183.66)	(62.47)	(41.29)	(94.28)	(444.52)	(793.72)	(6.34)
BLACK	-4.38	-4.3	-3.94	-5.07	-7.72	-9.9	-10.47	-6.22
	(.87)	(.8)	(.5)	(.53)	(.64)	(1.28)	(1.59)	(.54)
	***	***	***	***	***	***	***	***
HISPANIC	-1.24	-1.11	-2.01	-2.59	-2.65	-4.16	-3.32	-2.61
	(1.02)	(.71)	(.53)	(.51)	(.82)	(1.53)	(2.2)	(.52)
			***	***	***	***		***
ASIAN	-1.44	-2.27	-2.84	-3.36	-3.45	-3.74	-1.41	-2.92
	(1.06)	(.84)	(.92)	(.88)	(1.19)	(1.85)	(2.36)	(.79)
		***	***	***	***	**		***
ISOLATED	-4.1	-3.19	-4.39	-3.78	-5.26	-7.06	-8.81	-5.43
	(1.34)	(1.33)	(1.06)	(.96)	(1.11)	(1.95)	(2.46)	(.91)
		**	***	***	***	***	***	***

READING*SES	-.03 (.01) ***	-.04 (.01) ***	-.06 (.01) ***	-.05 (.01) ***	-.06 (.02) ***	-.07 (.02) ***	-.12 (.02) ***	-.05 (.01) ***
READING*GENDER	-.02 (.01) *	-.05 (.01) ***	-.08 (.01) ***	-.08 (.01) ***	-.08 (.02) ***	-.07 (.03) **	-.07 (.04) **	-.07 (.01) ***
READING*FORMER	.05 (.03) *	.03 (.02)	.02 (.03)	0 (.03)	-.02 (.04)	-.04 (.06)	.04 (.07)	0 (.02)
READING*ELL	.48 (43.85)	.45 (22.4)	.14 (6.19)	-.03 (4.87)	-.24 (1.41)	-.46 (55.96)	-.33 (88.08)	-.06 (.73)
READING*BLACK	0 (.02)	-.03 (.02)	-.01 (.02)	-.04 (.03)	-.09 (.04) **	-.06 (.06)	-.05 (.07)	-.04 (.02) *
READING*HISPANIC	-.09 (.02) ***	-.07 (.02) ***	-.05 (.03) **	-.02 (.03)	0 (.04)	0 (.05)	-.04 (.08)	-.04 (.02)
READING*ASIAN	-.03 (.03)	-.02 (.03)	-.05 (.03)	-.05 (.04)	0 (.05)	0 (.06)	.03 (.08)	-.02 (.03)
READING*ISOLATED	-.11 (.04) ***	-.1 (.03) ***	-.11 (.04) **	-.06 (.05)	-.03 (.06)	-.13 (.12)	-.11 (.17)	-.1 (.04) ***

Table 4.2b. Grade 3

	QRM							MRM
	.05	.1	.25	.5	.75	.9	.95	
Intercept	76.19	82.09	93.35	104.99	116.1	125.71	130.82	104.54
	(.61)	(.47)	(.49)	(.35)	(.39)	(.49)	(.4)	(.29)
	***	***	***	***	***	***	***	***
READING	.63	.63	.65	.62	.58	.51	.5	.6
	(.02)	(.02)	(.01)	(.01)	(.01)	(.02)	(.02)	(.01)
	***	***	***	***	***	***	***	***
SES	1.96	2.56	2.62	3.05	3.22	4.05	4.1	3.01
	(.55)	(.43)	(.36)	(.32)	(.4)	(.53)	(.39)	(.27)
	***	***	***	***	***	***	***	***
GENDER	-5	-4.95	-7.44	-8.22	-8.71	-8.68	-8.49	-7.67
	(.72)	(.52)	(.5)	(.46)	(.49)	(.56)	(.56)	(.36)
	***	***	***	***	***	***	***	***
FORMER	2.06	1.98	1.31	.78	-.02	.03	1.43	1.04
	(1.36)	(1.05)	(.94)	(.98)	(1)	(1.24)	(1.46)	(.71)
		*						
ELL	-7.92	-4.13	.23	-.63	5.27	3	2.03	-.06
	(4.46)	(5.2)	(4.42)	(4.01)	(5.06)	(4.62)	(6.67)	(2.74)
	*							
BLACK	-6.1	-5.74	-6.28	-7.71	-7.6	-4.72	-4.91	-6.75
	(1.16)	(1.11)	(.87)	(.79)	(.93)	(1.12)	(1.21)	(.68)
	***	***	***	***	***	***	***	***
HISPANIC	-1.7	-2.19	-3.13	-1.86	-.86	-.57	-1.09	-1.98
	(1.37)	(.9)	(.91)	(.87)	(1.01)	(.9)	(1.27)	(.63)
		**	***	**				***
ASIAN	.86	.4	.96	2.06	2.12	3.88	.63	1.49
	(1.88)	(1.39)	(1.63)	(1.21)	(1.23)	(1.5)	(1.61)	(.92)
				*	*	***		
ISOLATED	-1.57	-2.79	-3.02	-1.96	-1.85	-2.38	-1.79	-1.97
	(2.55)	(1.98)	(2.11)	(1.69)	(1.24)	(1.95)	(2.26)	(1.18)
								*

READING*SES	.06	.05	.02	0	-.02	-.05	-.07	-.01
	(.02)	(.01)	(.01)	(.01)	(.01)	(.02)	(.01)	(.01)
	***	***	**			***	***	
READING*GENDER	-.1	-.07	-.05	-.03	0	.04	.05	-.02
	(.03)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.01)
	***	***	***	*		*	**	
READING*FORMER	.04	.05	.01	-.05	0	.04	0	-.01
	(.05)	(.04)	(.03)	(.03)	(.04)	(.05)	(.06)	(.03)
READING*ELL	-.15	-.07	-.07	-.07	.06	-.03	-.08	-.06
	(.11)	(.11)	(.1)	(.09)	(.12)	(.12)	(.14)	(.07)
READING*BLACK	-.14	-.09	-.04	.01	.06	.11	.13	.01
	(.04)	(.04)	(.03)	(.03)	(.03)	(.04)	(.05)	(.03)
	***	**			**	***	**	
READING*HISPANIC	-.06	-.05	-.02	0	-.02	-.01	.02	0
	(.05)	(.04)	(.03)	(.03)	(.04)	(.04)	(.05)	(.02)
READING*ASIAN	.08	.02	.1	.14	.09	.07	.18	.1
	(.08)	(.06)	(.04)	(.04)	(.04)	(.06)	(.07)	(.03)
			***	***	**		***	***
READING*ISOLATED	-.13	-.12	-.07	.03	.05	.04	-.03	-.01
	(.1)	(.06)	(.05)	(.05)	(.04)	(.06)	(.09)	(.04)
		*						

Table 4.2c. Grade 5

	QRM							MRM
	.05	.1	.25	.5	.75	.9	.95	
Intercept	100.41	107.08	118.64	130.65	140.8	148.42	152.78	129.05
	(.61)	(.67)	(.43)	(.35)	(.34)	(.43)	(.4)	(.3)
	***	***	***	***	***	***	***	***
READING	.75	.74	.69	.58	.47	.39	.34	.58
	(.02)	(.02)	(.02)	(.01)	(.01)	(.02)	(.02)	(.01)
	***	***	***	***	***	***	***	***
SES	3.68	3.1	3.74	3.52	3.15	3.9	3.75	3.39
	(.48)	(.55)	(.42)	(.33)	(.29)	(.4)	(.35)	(.27)
	***	***	***	***	***	***	***	***
GENDER	-7.41	-6.87	-7.31	-7.5	-7.23	-6.68	-6.74	-7.3
	(.72)	(.76)	(.53)	(.43)	(.37)	(.5)	(.47)	(.35)
	***	***	***	***	***	***	***	***
FORMER	.24	1.08	2.17	1.3	1.05	1.05	.81	1.56
	(1.71)	(1.67)	(1.22)	(.83)	(.94)	(1.01)	(1.22)	(.72)
			*					**
ELL	-2.18	1.91	9.06	6.93	6.25	3.28	8.57	5.97
	(6.29)	(7.07)	(4.06)	(3.43)	(2.77)	(5.17)	(6.01)	(2.34)
			**	**	**			**
BLACK	-7.07	-5.42	-6.41	-7.87	-7.65	-5.82	-3.81	-6.85
	(1.82)	(1.3)	(.87)	(1.07)	(.84)	(1.26)	(1.31)	(.68)
	***	***	***	***	***	***	***	***
HISPANIC	-1.94	-2.51	-1.46	-1.06	-1.71	-1.66	-1.33	-1.55
	(1.51)	(1.52)	(1.04)	(.75)	(.76)	(.89)	(1.13)	(.63)
		*			**	*		**
ASIAN	2.81	2.5	5.48	6.29	4.03	3.94	4.52	4.63
	(2.52)	(2.13)	(1.44)	(1.24)	(1.1)	(1.27)	(1.17)	(.91)
			***	***	***	***	***	***
ISOLATED	1.6	-.37	-1.32	-2.41	-4.18	-1	-.63	-1.64
	(3.58)	(1.85)	(1.54)	(1.26)	(1.53)	(1.88)	(3.13)	(1.16)
				*	***			

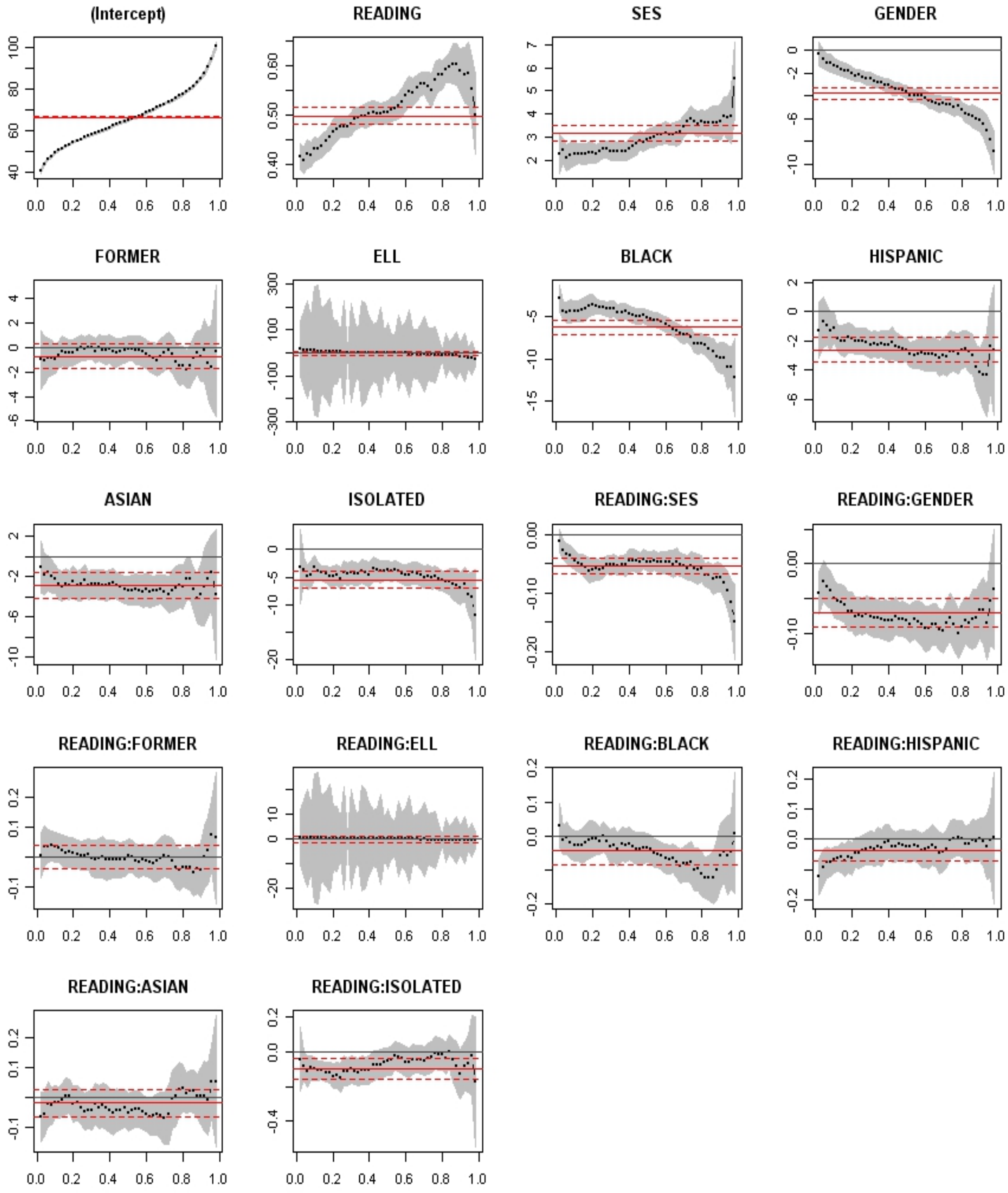
READING*SES	.01 (.02)	.03 (.02) *	-.03 (.01) **	-.05 (.01) ***	-.05 (.01) ***	-.07 (.01) ***	-.07 (.01) ***	-.04 (.01) ***
READING*GENDER	0 (.03)	.01 (.02)	.03 (.02)	.05 (.01) ***	.07 (.01) ***	.07 (.02) ***	.09 (.02) ***	.04 (.01) ***
READING*FORMER	-.02 (.06)	-.03 (.05)	.04 (.04)	0 (.03)	0 (.04)	0 (.04)	.03 (.06)	.01 (.03)
READING*ELL	-.15 (.16)	-.05 (.15)	.16 (.11)	.1 (.08)	.15 (.09) *	.02 (.14)	.13 (.17)	.09 (.06)
READING*BLACK	-.15 (.06) **	-.03 (.04)	-.03 (.03)	.08 (.03) **	.16 (.03) ***	.13 (.05) **	.18 (.06) ***	.06 (.02)
READING*HISPANIC	.04 (.07)	.03 (.05)	-.03 (.04)	.01 (.03)	.04 (.03)	.06 (.04)	.06 (.06)	.02 (.03)
READING*ASIAN	.12 (.07) *	.1 (.07)	-.06 (.05)	-.03 (.04)	.05 (.04)	.07 (.06)	.01 (.07)	.03 (.04)
READING*ISOLATED	-.15 (.12)	-.05 (.05)	-.09 (.04) **	0 (.04)	.06 (.05)	.11 (.09)	.05 (.12)	.01 (.04)

Table 4.2d. Grade 8

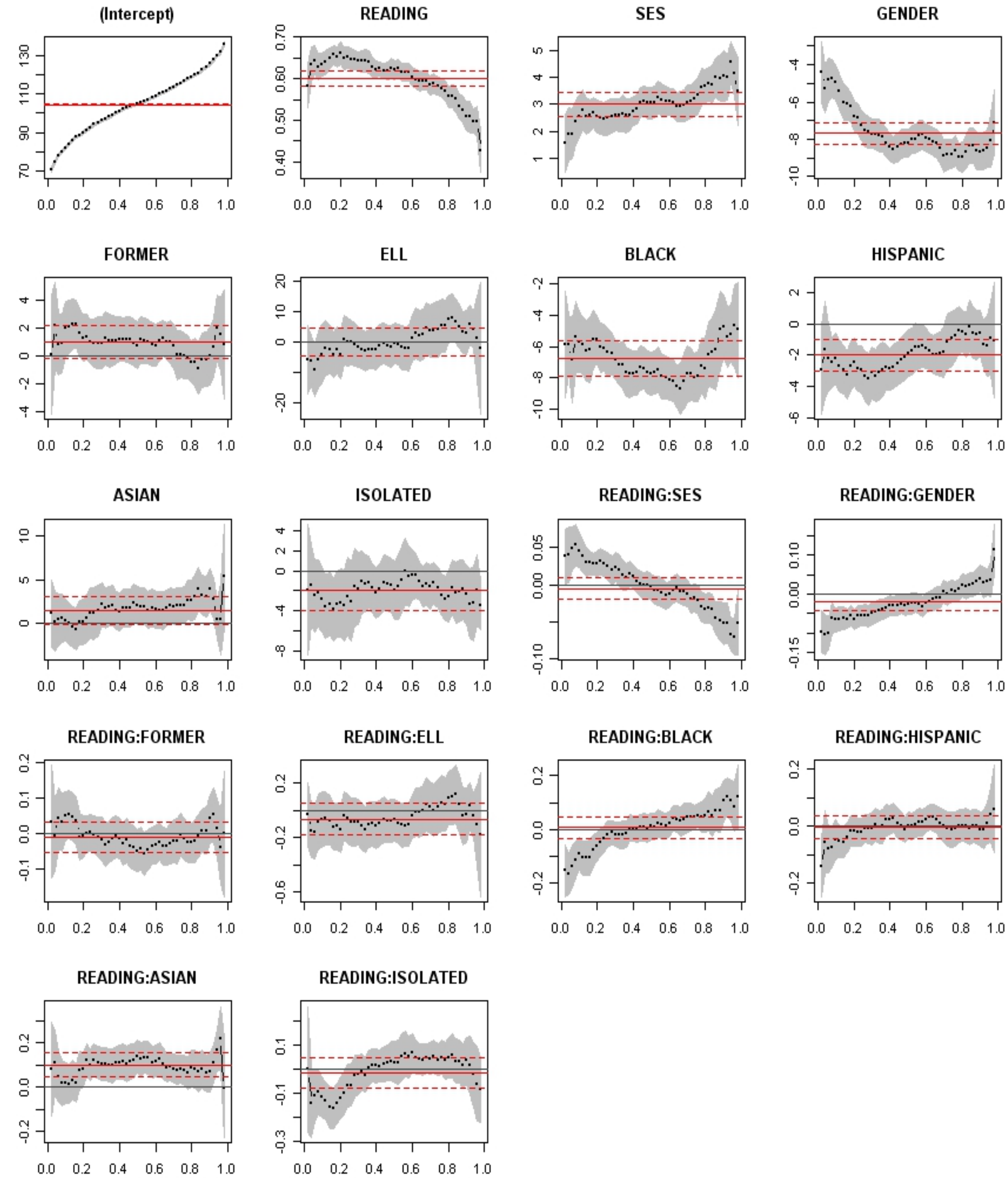
	QRM							MRM
	.05	.1	.25	.5	.75	.9	.95	
Intercept	120.27	126.48	137.37	147.42	156.18	162.35	165.42	145.81
	(.66)	(.53)	(.36)	(.32)	(.26)	(.22)	(.25)	(.27)
	***	***	***	***	***	***	***	***
READING	.74	.73	.62	.53	.39	.28	.2	.51
	(.03)	(.02)	(.02)	(.01)	(.01)	(.01)	(.01)	(.01)
	***	***	***	***	***	***	***	***
SES	3.48	3.69	3.59	2.94	2.87	2.28	1.95	3.11
	(.59)	(.42)	(.31)	(.27)	(.21)	(.22)	(.24)	(.24)
	***	***	***	***	***	***	***	***
GENDER	-5.1	-5.53	-6.42	-5.31	-4.09	-3.27	-3.25	-4.98
	(.82)	(.61)	(.46)	(.37)	(.31)	(.31)	(.3)	(.32)
	***	***	***	***	***	***	***	***
FORMER	-.29	-1.07	-2.87	-1.23	-.16	.26	-.23	-.72
	(1.62)	(1.52)	(1.01)	(.72)	(.6)	(.61)	(.65)	(.64)
			***	*				
ELL	-3.89	-.16	.93	-1.04	0	-1.2	-2.66	-.41
	(6.99)	(4.84)	(2.76)	(1.94)	(2.7)	(3.35)	(3.89)	(1.81)
BLACK	-5.03	-4.66	-4.42	-5.15	-4.73	-3.81	-2.74	-4.64
	(1.54)	(1.32)	(1.08)	(.9)	(.73)	(.85)	(.92)	(.65)
	***	***	***	***	***	***	***	***
HISPANIC	-1.52	.51	.96	.37	-.43	-.04	.12	.13
	(1.34)	(1.45)	(.89)	(.7)	(.58)	(.54)	(.55)	(.56)
ASIAN	-.52	3.63	4.68	4.81	3.63	3.1	3.19	3.73
	(2.8)	(2.26)	(1.27)	(.96)	(.76)	(.65)	(.85)	(.84)
			***	***	***	***	***	***
ISOLATED	-2.59	-.8	-2.71	-.78	-.66	-.18	-.38	-.86
	(3.13)	(2.16)	(1.79)	(1.66)	(.93)	(1.53)	(1.6)	(1.04)

READING*SES	.04 (.02) **	.03 (.02) **	.01 (.01)	-.03 (.01) ***	-.06 (.01) ***	-.07 (.01) ***	-.06 (.01) ***	-.03 (.01) ***
READING*GENDER	-.06 (.03) **	-.06 (.03) **	-.01 (.02)	.05 (.01) ***	.07 (.01) ***	.07 (.01) ***	.08 (.01) ***	.02 (.01) *
READING*FORMER	.05 (.05)	.02 (.04)	.03 (.03)	.08 (.03) ***	.02 (.02)	.01 (.03)	.03 (.03)	.04 (.02)
READING*ELL	-.11 (.14)	.03 (.1)	.07 (.05)	.03 (.06)	-.02 (.08)	-.12 (.11)	-.06 (.12)	.01 (.04)
READING*BLACK	-.12 (.04) ***	-.11 (.05) **	-.03 (.03)	-.01 (.03)	.01 (.03)	.1 (.03) ***	.09 (.03) ***	-.02 (.02)
READING*HISPANIC	-.07 (.05)	-.06 (.04)	-.03 (.03)	-.05 (.03) *	.02 (.02)	.03 (.03)	.01 (.03)	-.02 (.02)
READING*ASIAN	.14 (.1)	.04 (.09)	.02 (.04)	-.03 (.04)	-.05 (.03)	-.07 (.03) **	-.09 (.04) **	-.01 (.03)
READING*ISOLATED	-.08 (.14)	0 (.07)	.09 (.06)	.06 (.05)	.14 (.03) ***	.07 (.08)	.07 (.09)	.09 (.03)

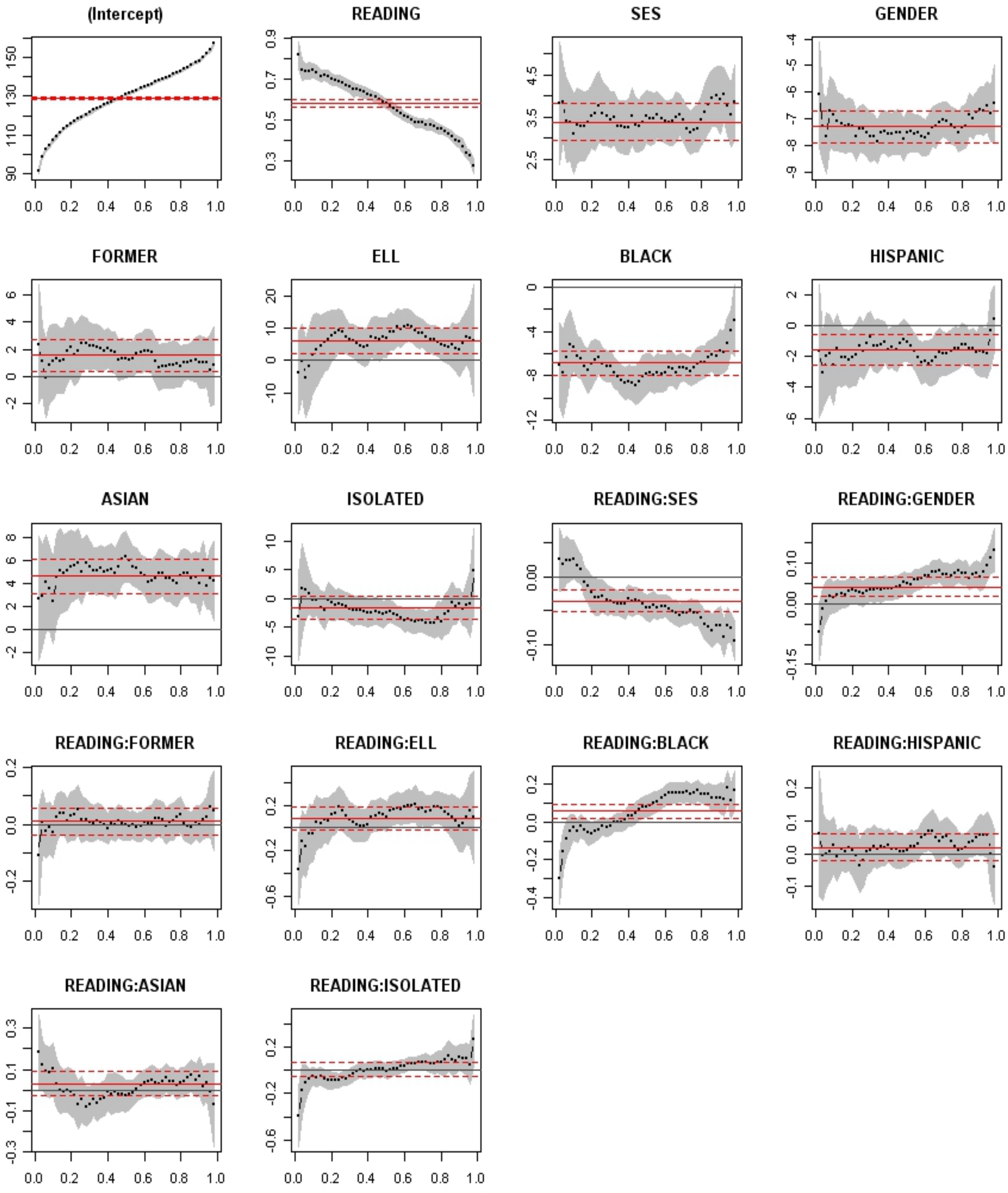
APPENDIX E: FULL QUANTILE PROCESS PLOT BY GRADE



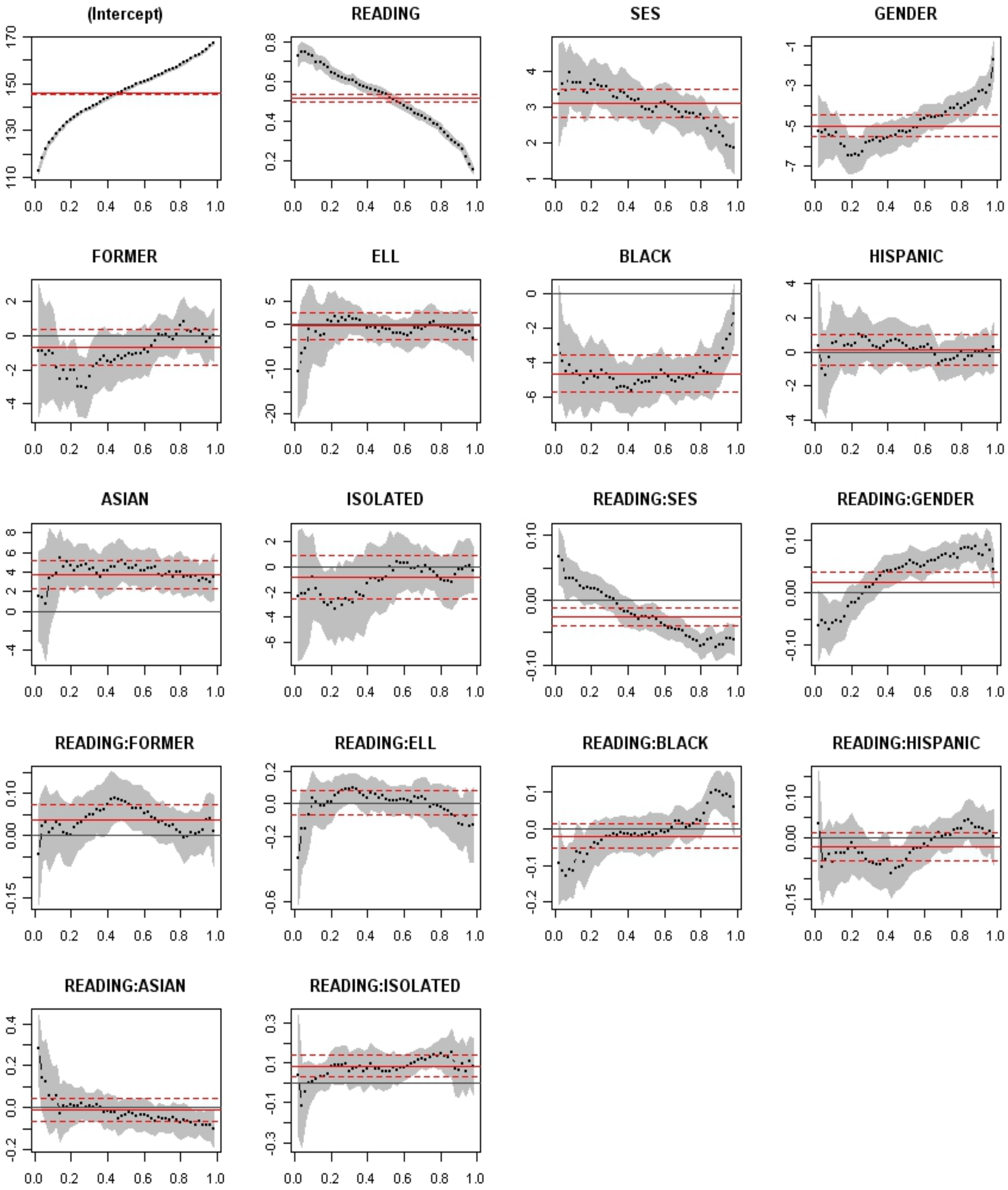
Grade 1


(Intercept)
70
90
110
130
READING
0.40
0.50
0.60
0.70
SES
1
2
3
4
5
GENDER
-10
-8
-6
-4
FORMER
-4
-2
0
2
4
ELL
-20
0
10
20
BLACK
-10
-8
-6
-4
-2
HISPANIC
-6
-4
-2
0
2
ASIAN
0
5
10
ISOLATED
-8
-4
0
2
4
READING:SES
-0.10
0.00
0.05
READING:GENDER
-0.15
0.00
0.10
READING:FORMER
-0.1
0.0
0.1
0.2
READING:ELL
-0.6
-0.2
0.2
READING:BLACK
-0.2
0.0
0.2
READING:HISPANIC
-0.2
0.0
0.2
READING:ASIAN
-0.2
0.0
0.2
READING:ISOLATED
-0.3
-0.1
0.1
0.0
0.2
0.4
0.6
0.8
1.0

Grade 3


(Intercept)
90
110
130
150
READING
0.3
0.5
0.7
0.9
SES
2.5
3.5
4.5
GENDER
-9
-8
-7
-6
-5
-4
FORMER
-2
0
2
4
6
ELL
-10
0
10
20
BLACK
-12
-8
-4
0
HISPANIC
-6
-4
-2
0
2
ASIAN
-2
0
2
4
6
8
ISOLATED
-10
-5
0
5
10
READING:SES
-0.10
0.00
READING:GENDER
-0.15
0.00
0.10
READING:FORMER
-0.2
0.0
0.2
READING:ELL
-0.6
-0.2
0.2
READING:BLACK
-0.4
-0.2
0.0
0.2
READING:HISPANIC
-0.1
0.0
0.1
0.2
READING:ASIAN
-0.3
-0.1
0.1
0.3
READING:ISOLATED
-0.6
-0.2
0.2
0.0
0.2
0.4
0.6
0.8
1.0

Grade 5


(Intercept)
READING
SES
GENDER
FORMER
ELL
BLACK
HISPANIC
ASIAN
ISOLATED
READING:SES
READING:GENDER
READING:FORMER
READING:ELL
READING:BLACK
READING:HISPANIC
READING:ASIAN
READING:ISOLATED
110
130
150
170
0.2
0.4
0.6
0.8
1
2
3
4
-7
-5
-3
-1
-4
-2
0
2
-20
-10
0
5
-6
-4
-2
0
-4
-2
0
2
4
-4
0
2
4
6
8
-6
-4
-2
0
2
-0.10
0.00
0.10
-0.10
0.00
0.10
-0.15
0.00
0.10
-0.6
-0.2
0.0
0.2
-0.2
-0.1
0.0
0.1
-0.15
0.00
0.10
-0.2
0.0
0.2
0.4
-0.3
-0.1
0.1
0.3
0.0
0.2
0.4
0.6
0.8
1.0

Grade 8

APPENDIX F: KHMALADEZ TEST AT GRADE LEVEL

	Grade 1		Grade 3		Grade 5		Grade 8	
	L Shift	LS Shift	L Shift	LS Shift	L Shift	LS Shift	L Shift	LS Shift
READING	6.64 ***	1.11	1.67	.52	.90	.89	6.63 ***	3.64 ***
SES	1.58	1.26	.53	.81	.22	.41	1.27	.43
GENDER	2.38 **	.33	.93	.51	.23	.78	4.37 ***	1.96
FORMER	.53	.48	1.87	.84	.31	.72	1.53	1.45
ELL	4.97 ***	.46	.37	.44	.22	.88	.68	1.24
BLACK	3.98 ***	.64	1.18	1.29	.45	1.27	1.41	1.18
HISPANIC	1.52	.59	1.60	.57	.37	.62	.83	1.42
ASIAN	.46	.33	.45	.54	1.40	.94	3.26 ***	1.00
ISOLATED	1.00	.66	1.31	.37	.62	.36	1.92	.92
Overall	23.99 ***	11.88 ***	7.91	4.50	3.71	6.29	48.12 ***	18.97 ***

Note:
L Shift: Location shift only hypothesis
LS Shift: Location-scale shift hypothesis
*** significant at or below .01
** significant at or below .05