观海文丛——
华东师范大学外语
学院学者文库

Differential Language Impact in Math Assessment

语言能力在数学测试中的差异性影响

陈　芳·著

南開大學出版社

观海文丛——华东师范大学外语学院学者文库

语言能力在数学测试中的差异性影响

Differential Language Impact in Math Assessment

陈 芳 著

南开大学出版社
天 津

图书在版编目(CIP)数据

语言能力在数学测试中的差异性影响 / 陈芳著. —天津：南开大学出版社，2014.6

（观海文丛.华东师范大学外语学院学者文库）

ISBN 978-7-310-04500-6

Ⅰ.①语… Ⅱ.①陈… Ⅲ.①语言能力—影响—数学—教学—教学研究 Ⅳ.①H0②01—4

中国版本图书馆 CIP 数据核字(2014)第 101989 号

南开大学出版社出版发行

出版人：孙克强

地址：天津市南开区卫津路 94 号 邮政编码：300071

营销部电话：(022)23508339 23500755

营销部传真：(022)23508542 邮购部电话：(022)23502200

天津泰宇印务有限公司印刷

全国各地新华书店经销

2014 年 6 月第 1 版 2014 年 6 月第 1 次印刷

260×185 毫米 16 开本 7.5 印张 2 插页 189 千字

定价：18.00 元

如遇图书印装质量问题，请与本社营销部联系调换，电话：(022)23507125

序

语言作为思维的载体，对于学科学习的影响不容忽视。大量研究表明学科成绩与语言能力密切相关，语言能力低的人（不论是母语使用者，还是二语习得者）也通常表现为学科成绩较差。对于这个现象有不同的看法，有的学者认为这是总的智商的差异造成的，有的则认为这是由语言在教学中的媒介作用决定的。教育界和测量界的人更愿意接受后者的解释，因为当差异被归于基因时，也就意味着我们无法介入。而教育的目的和功能之一就是介入教与学的过程以最大限度地开发所有人的潜能，提供公平的发展机会。从这个角度来看，可以控制的变量才是值得研究的变量。语言能力的培养是可以控制的变量，课程体系中的语言域（语言文学）和其他学科语言域（学术语言）的差别和联系也可以研究，因而研究语言能力与其他学科的关系也就具有可操作性和实际指导意义。

语言能力对于学科成就的影响在美国这样的移民国家里更倍受关注。英语（作为第二语言的）习得者（English Language Learners, ELLs）的多项学科成绩普遍低于英语为母语的同龄人，这给美国的学科教师们和教育管理层带来了很大的压力和挑战。虽然由于语言能力的局限，英语习得者的学业受到了影响，然而这些影响到底有多大，这些影响的强度是不是一个固定值？本书以数学学科为例探讨了语言能力与数学成绩的长期的、变动的关系。本案例使用了美国一项追踪研究的数据分析这二者在美国教学环境下的关系，但其研究发现对于国内双语学校教学研究的价值也是显而易见的。本文的方法放在国内也可用于研究少数民族学生在以普通话为教学和考试语言的各类学校中的学业成就差异。

探讨变量之间关系的一个常规的统计模型是回归。传统回归是以平均值为研究对象的，即回归分析中的参数反映的都是就平均而言的变量之间的关系。然而变量之间的关系不一定是固定的。当我们试图区分不同的能力阶层中语言和学科能力之间的差异性关系时，我们的研究目标就是一个变动的概念。本研究使用了一种与此目的相匹配的回归模型，即分位数回归，来检验语言能力与数学成绩之间的差异性关系。分位数回归在自然科学、经济学等领域已经广泛使用。国际上，教育界有少量文章运用了分位数回归，但一般是运用在探讨教育投资与收益的经济学问题上，然而这个模型对于教育测量和评价领域是个极富价值的工具。美国的“不让一个孩子掉队”（No Child Left Behind）法律要求各州使用科学的方法展示和评估本校学生的学业进步，以分位数回归为基础的“学生成长百分点”（Student Growth Percentile, SGP）模型就以其特有的分层分析的优势脱颖而出，得到美国教育部认可，成为符合政府要求的绩效评估模型之一。

语言对于学业成就有多个方面的影响，比如教、学和评估。本书从教育测量的角度出发，因此主要关注测试中的语言影响。本研究首次将分位数回归运用于国际语言测试领域关注的一个研究课题中，从教育测量的角度探讨测试语言对于数学成绩的区分性影响。本研究既用实证案例详细讲解了分位数的方法，突出对比了以往研究中的方法论上的局限性，同时验证了一些语言测试领域的推测和早期结论，为语言测试和学科测试与教学提供了一

些改进的依据。此外，分位数回归模型通常同时运行若干个子模型，因此数据的呈现和解读也比较复杂。本研究用大量图表直观呈现数据结果，并针对分位数回归的特点设计如图4.6的方式辅助判断I型和II型错误的可能性。本研究也因此尝试缓期决策的建模程序，以稳定的规律而不是单个模型的结果来得出结论。

总而言之，本研究的目的是抛砖引玉，希望通过这个研究能激发更多人运用这个新模型来更深入地研究教育领域，特别是测量和评估领域的类似问题，增长我们对于众多现象的更全面的理解。

Abstract

New models are commonly developed to overcome certain limitations of other ones. Quantile regression is introduced in this book because it can provide the information which 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 the research in application of quantile regression aims 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. Generally speaking, the math gap between different groups on the basis of language proficiency levels or ethnicity varies between students with different math ability. 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 Non-ELLs. Model-building process suggests that language demand in tests may have contributed to the big achievement gap between ELLs and Non-ELLs in USA. This confirms previous conclusions on the necessity of accommodations and the importance of appropriate type of accommodations in academic tests administered in English. Results provide evidence supporting the long-lasting effect of language on math achievement. However, the long term relationship between language and math is not a constant; rather, it seems to decrease after controlling for other relevant factors in education. Differential effects of several background variables are also detected.

Implication of the research results for teaching and testing is discussed. Limitations of the quantile regression technique are revealed as a guide for future research.

南开大学出版社网址：http://www.nkup.com.cn

投稿电话及邮箱：022-23504636　QQ：1760493289
QQ：2046170045(对外合作)
邮购部：022-23507092
发行部：022-23508339　Fax：022-23508542

南开教育云：http://www.nkcloud.org

App：南开书店 app

南开教育云由南开大学出版社、国家数字出版基地、天津市多媒体教育技术研究会共同开发，主要包括数字出版、数字书店、数字图书馆、数字课堂及数字虚拟校园等内容平台。数字书店提供图书、电子音像产品的在线销售；虚拟校园提供 360 校园实景；数字课堂提供网络多媒体课程及课件、远程双向互动教室和网络会议系统。在线购书可免费使用学习平台，视频教室等扩展功能。

Table of Contents

List of Tables and Figures

Chapter I Introduction

The first sentence in Chapter 9 of the fourth edition of *Standards for Educational and Psychological Testing* (hereinafter, "the Standards", AERA, APA & NCME, 1999) clearly states that, "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 of 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 such as science or social studies to detect language impact in those subject area assessments. The relationship between language and math has been studied for a long time and it is 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. In addition to that, longitudinal data used in this research will shed light onto the long-term language effect on academic performance that is of interest to the language testing field yet has not been well represented in past research.

1.1 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. The heated issue is also reflected in the discussion on the achievement gaps between native English speakers (Non English Language Learners or 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: 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 in the classroom; 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 analyses on 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 ELLs 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 up the grades. There are several speculations on this discrepancy. For example, the ELL groups were defined according to different criterion in these studies: not only the number of categories but also the sample composition differ in each study. The math measurements adopted in each study also differed from each other, which implies inequality of test specification, psychometric quality and language requirement of the various instruments among studies. Other background factors may have also interfered with the relationship between language and math. All these naturally led to different research results. Above all these, the relationship between language and math may not be static but change indeed across grades as well as within each grade. More importantly, few studies directly controlled for the language proficiency of individual students before comparing their math scores, which means the impact of language in math assessment for individual students was ignored, resulting in an inaccurate representation of the overall math achievement gap between ELLs and Non-ELLs as two groups.

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 issue of sample difference. A quantile regression methodology that models more than the central tendency can handle the differential language effect at various math ability levels. Adding the language proficiency covariate solves the last limitation mentioned in the paragraph above.

1.2 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 for all students with low math ability. Therefore, language support should be given to all students be them ELLs or not. This information may lead to alternative resource allocation policies in schools. 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, analysis should start with a purer or adjusted math component. For this purpose, it is suggested that the language variance should 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 many advantages to be discussed 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?

1.3 Theoretical Background

1.3.1 English Language Learners and the Inequality in Educaiton

"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; U.S. 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 causes 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).

1.3.2 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 has been 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 and 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, that is, listening, speaking and writing.

1.3.3 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 et al., 2000) and the meaning of standardized test 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). For example, 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 the performance of students 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 achievement 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" (p.101).

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 can make 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 to "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), aim to use innovative computer-based items with minimal language requirement to assess ELLs (Kopriva et al., 2009). However, to ensure the 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).

1.4 Chapter Summary

Language proficiency affects content knowledge learning and assessment especially for ELLs. While 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 among students within and across grades. To better explore