Title
Essays on Education Policy /

Permalink
https://escholarship.org/uc/item/72q033vr

Author
Zieleniak, Jedrzej

Publication Date
2014

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Essays on Education Policy**

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Economics

by

Jedrzej Zieleniak

Committee in charge:

Professor Julian R. Betts, Chair
Professor Craig T. McIntosh, Co-Chair
Professor Prashant Bharadwaj
Professor Amanda Datnow
Professor Karthik Muralidharan

2014
The Dissertation of Jedrzej Zieleniak is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2014
# TABLE OF CONTENTS

Signature Page ................................................................. iii  
Table of Contents ............................................................... iv  
List of Figures ................................................................ vii  
List of Tables ................................................................ viii  
Acknowledgements ............................................................... x  
Vita ............................................................................... xi  
Abstract of the Dissertation .................................................... xii  

Chapter 1 Jumping Through Hoops: Student Achievement and the California High School Exit Exam ........................................... 1  
1.1 Abstract .................................................................. 2  
1.2 Introduction .............................................................. 3  
1.3 Institutional Context ...................................................... 7  
  1.3.1 California High School Exit Exam .................. 7  
  1.3.2 Student Testing and Reporting (STAR) ........ 10  
1.4 Data .................................................................. 11  
1.5 Empirical Framework .................................................... 13  
  1.5.1 Student Achievement While in High School ...... 14  
  1.5.2 Effects of the CAHSEE on Graduation Rates .. 17  
1.6 Results .................................................................. 19  
  1.6.1 Student Achievement on Standardized Tests ... 19  
  1.6.2 Exit Exam and Graduation ............................ 24  
  1.6.3 Exit Exam and Math Course Performance ...... 30  
1.7 Conclusions ................................................................. 35  
1.8 Tables .................................................................. 36  
1.9 Figures .................................................................. 45  
1.10 Model of Exit Exams and Achievement ...................... 49  
  1.10.1 Student Achievement and Testing ............... 49  
  1.10.2 Exit Exams and Distribution of Achievement .. 51  
  1.10.3 Exit Exams and Incentives to Learn .............. 53  
1.11 Acknowledgments ........................................................ 54  

Chapter 2 Effects of Test Failure on Academic Achievement in the Context of Student Accountability: A Regression Discontinuity Analysis. 55  
2.1 Abstract .................................................................. 56  
2.2 Introduction .............................................................. 57
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3 The California High School Exit Exam</td>
<td>61</td>
</tr>
<tr>
<td>2.4 Conceptual Framework and Data</td>
<td>63</td>
</tr>
<tr>
<td>2.4.1 RD Design for Student Achievement</td>
<td>64</td>
</tr>
<tr>
<td>2.4.2 Empirical Design</td>
<td>66</td>
</tr>
<tr>
<td>2.4.3 Bandwidth Selection</td>
<td>67</td>
</tr>
<tr>
<td>2.4.4 Data</td>
<td>68</td>
</tr>
<tr>
<td>2.5 Empirical Results</td>
<td>71</td>
</tr>
<tr>
<td>2.5.1 Remedial Courses and Academic Trajectories</td>
<td>73</td>
</tr>
<tr>
<td>2.5.2 Test Failure and Performance in High School</td>
<td>76</td>
</tr>
<tr>
<td>2.5.3 CAHSEE Failure, Dropouts and Graduation</td>
<td>78</td>
</tr>
<tr>
<td>2.5.4 Postsecondary Outcomes</td>
<td>79</td>
</tr>
<tr>
<td>2.6 Conclusions</td>
<td>81</td>
</tr>
<tr>
<td>2.7 Tables</td>
<td>84</td>
</tr>
<tr>
<td>2.8 Figures</td>
<td>87</td>
</tr>
<tr>
<td>2.9 Validity of the RD Framework</td>
<td>89</td>
</tr>
<tr>
<td>2.10 Supplementary Tables and Figures</td>
<td>92</td>
</tr>
<tr>
<td>2.11 Acknowledgements</td>
<td>100</td>
</tr>
<tr>
<td>Chapter 3 Chasing the Syllabus: Measuring Learning Trajectories in Developing Countries with Longitudinal Data and Item Response Theory</td>
<td>101</td>
</tr>
<tr>
<td>3.1 Abstract</td>
<td>102</td>
</tr>
<tr>
<td>3.2 Introduction</td>
<td>103</td>
</tr>
<tr>
<td>3.3 Estimating Learning Trajectories</td>
<td>108</td>
</tr>
<tr>
<td>3.3.1 Methodological Issues</td>
<td>108</td>
</tr>
<tr>
<td>3.3.2 Item-Response Theory (IRT) models</td>
<td>109</td>
</tr>
<tr>
<td>3.3.3 Vertically-Scaled IRT models</td>
<td>110</td>
</tr>
<tr>
<td>3.4 Validating the vertically-scaled IRT model</td>
<td>112</td>
</tr>
<tr>
<td>3.4.1 Context and Data</td>
<td>112</td>
</tr>
<tr>
<td>3.4.2 Linking, Equating, and Validation Design</td>
<td>114</td>
</tr>
<tr>
<td>3.4.3 Results: Validation of vertically-scaled IRT in cross-sectional data</td>
<td>118</td>
</tr>
<tr>
<td>3.4.4 Results: Validation of vertically-scaled IRT in longitudinal data</td>
<td>120</td>
</tr>
<tr>
<td>3.5 Learning Trajectories</td>
<td>121</td>
</tr>
<tr>
<td>3.5.1 Levels</td>
<td>121</td>
</tr>
<tr>
<td>3.5.2 Inequality</td>
<td>126</td>
</tr>
<tr>
<td>3.5.3 Gaps</td>
<td>128</td>
</tr>
<tr>
<td>3.5.4 Flattening/learning to read.</td>
<td>130</td>
</tr>
<tr>
<td>3.5.5 Productivity of the school system</td>
<td>131</td>
</tr>
<tr>
<td>3.5.6 Progress Over Time</td>
<td>132</td>
</tr>
<tr>
<td>3.6 Conclusions</td>
<td>133</td>
</tr>
<tr>
<td>3.7 Item Response Theory- Appendix 1</td>
<td>135</td>
</tr>
<tr>
<td>Section</td>
<td>Page Number</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>3.8 Tables</td>
<td>138</td>
</tr>
<tr>
<td>3.9 Figures</td>
<td>144</td>
</tr>
<tr>
<td>3.10 Acknowledgements</td>
<td>149</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Bibliography</td>
<td>150</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Correlations between SAT9 and CST Scores in Mathematics</td>
<td>45</td>
</tr>
<tr>
<td>1.2</td>
<td>Correlations between SAT9 and CST Scores in ELA</td>
<td>45</td>
</tr>
<tr>
<td>1.3</td>
<td>Student Scores and GPA in Grades 8 and 10 in Math and English by School Year</td>
<td>46</td>
</tr>
<tr>
<td>1.4</td>
<td>Exit Exam Effects on Test Scores by Decile of Achievement</td>
<td>46</td>
</tr>
<tr>
<td>1.5</td>
<td>Exit Exam Effects on GPA by Decile of Achievement</td>
<td>47</td>
</tr>
<tr>
<td>1.6</td>
<td>Dropout Rates by Decile and Exam Regime</td>
<td>47</td>
</tr>
<tr>
<td>1.7</td>
<td>On-Time Graduation Rates by Decile and Exam Regime</td>
<td>48</td>
</tr>
<tr>
<td>1.8</td>
<td>Graduation Rates by Decile and Exam Regime</td>
<td>48</td>
</tr>
<tr>
<td>2.1</td>
<td>The Effects of Failing CAHSEE ELA in Grade 10 on the Probability of Enrollment in an ELA Preparatory Course in Grade 11</td>
<td>87</td>
</tr>
<tr>
<td>2.2</td>
<td>The Effects of Failing CAHSEE Math in Grade 10 on the Probability of Enrollment in an Math Preparatory Course in Grade 11</td>
<td>87</td>
</tr>
<tr>
<td>2.3</td>
<td>The Effects of Failing CAHSEE ELA in Grade 10 on Student Outcomes in Grade 11</td>
<td>88</td>
</tr>
<tr>
<td>2.4</td>
<td>The Effects of Failing CAHSEE Math in Grade 10 on Student Outcomes in Grade 11</td>
<td>88</td>
</tr>
<tr>
<td>A.2.1</td>
<td>Histogram Plots of the Scaled CAHSEE ELA and Scaled CAHSEE Math Scores in Grade 10</td>
<td>92</td>
</tr>
<tr>
<td>3.1</td>
<td>Graphical Representation of a Hypothetical Educational System</td>
<td>144</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of Actual and Predicted Responses to Overlapping Common Questions in the Cross Sectional Data</td>
<td>144</td>
</tr>
<tr>
<td>3.3</td>
<td>Comparison of Actual and Predicted Responses in the Longitudinal Data</td>
<td>145</td>
</tr>
<tr>
<td>3.4</td>
<td>IRT Model Prediction Errors in the Longitudinal Data</td>
<td>145</td>
</tr>
<tr>
<td>3.5</td>
<td>Evolution of the EAP Ability Scores in the 5-Year Cohort</td>
<td>146</td>
</tr>
<tr>
<td>3.6</td>
<td>Predicted Probability of a Correct Answer on a Grade 1 Question in the 5-Year Cohort</td>
<td>146</td>
</tr>
<tr>
<td>3.7</td>
<td>Predicted Probability of a Correct Answer on a Grade 2 Question in the 4- and 5-Year Cohorts</td>
<td>147</td>
</tr>
<tr>
<td>3.8</td>
<td>Predicted Probability of a Correct Answer on a Grade 3 Question in the 3-, 4- and 5-Year Cohorts</td>
<td>147</td>
</tr>
<tr>
<td>3.9</td>
<td>Predicted Probability of a Correct Answer on a Grade 3 Question in the 2-, 3-, 4- and 5-Year Cohorts</td>
<td>148</td>
</tr>
<tr>
<td>3.10</td>
<td>Evolution of the EAP Ability Scores in the Longitudinal Data</td>
<td>148</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

| Table 1.1: | Timeline of the CAHSEE by Grade Level and School Year | 37 |
| Table 1.2: | Student Characteristics by Exam Type, Grade 6-12 | 37 |
| Table 1.3: | Testing Parallel Trends, Using 1997-1998 to 2000-2001 Data | 38 |
| Table 1.4: | Triple Difference Estimate of the Impact of the CAHSEE on Low-Stakes Test Scores | 39 |
| Table 1.5: | Triple Difference Estimate of the Impact of the CAHSEE on Student Grades | 40 |
| Table 1.6: | Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of Dropping Out by Decile of Initial Achievement (dp/dx) | 41 |
| Table 1.7: | Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of On-Time Graduation by Decile of Initial Achievement (dp/dx) | 41 |
| Table 1.8: | Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of Graduation by Decile of Initial Achievement (dp/dx) | 42 |
| Table 1.9: | Triple Difference Estimate of the Impact of the CAHSEE on Math Scores by Math Course Taken | 43 |
| Table 1.10: | Triple Difference Estimate of the Impact of the CAHSEE on Math Grades by Math Course Taken | 44 |
| Table 2.1: | Means of Baseline Student Characteristics in Grade 9 by Grade 10 CAHSEE Performance Subgroup | 84 |
| Table 2.2: | Tests of Equality of Means of Grade 9 Characteristics by Grade 10 CAHSEE Performance Subgroup | 84 |
| Table 2.3: | Effects of Failing CAHSEE in Grade 10 on the Probability of Enrolling in a CAHSEE ELA Preparatory Class in Grade 11 | 85 |
| Table 2.4: | Effects of Failing CAHSEE in Grade 10 on Student Achievement on ELA Low-Stakes Tests in Grade 11 | 85 |
| Table 2.5: | Effects of Failing CAHSEE in Grade 10 on High School Completion Outcomes | 86 |
| Table 2.6: | Effects of Failing CAHSEE in Grade 10 on Post-Secondary Academic Outcomes | 86 |
| Table A.2.1: | Effects of Failing CAHSEE in Grade 10 on the Probability of Enrolling in CAHSEE Math Preparatory Class in Grade 11 | 92 |
| Table A.2.2: | Effects of Failing CAHSEE in Grade 10 on Student Achievement on ELA Low-Stakes Tests in Grade 11 | 93 |
| Table A.2.3: | Effects of Failing CAHSEE in Grade 10 on Student Achievement on Math Low-Stakes Tests in Grade 11 | 93 |
| Table A.2.4: | Effects of Failing CAHSEE in Grade 10 on Grade Point Average in Grade 11 | 94 |
Table A.2.5: Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of Dropping Out by Decile of Initial Achievement (dp/dx) .......................................................... 94
Table A.2.6: Effects of Failing CAHSEE in Grade 10 on the Probability of Graduating from High School 4 Years after Enrolling in Grade 9 ......................................................... 95
Table A.2.7: Effects of Failing CAHSEE in Grade 10 on the Probability of Graduating from High School Within 6 Years of Enrolling in Grade 9 .................................................. 95
Table A.2.8: The Effect of Failing CAHSEE in Grade 10 on the Probability of Enrolling in a 2-Year College After High School ......................................................... 96
Table A.2.9: The Effect of Failing CAHSEE in Grade 10 on the Probability of Enrolling in a 4-Year College After High School ......................................................... 96
Table A.2.10: The Effect of Failing CAHSEE in Grade 10 on the Probability of Transferring From a 2-Year to a 4-Year College ......................................................... 97
Table A.2.11: Robustness Checks: Manipulation of the Running Variable ......................................................... 97
Table A.2.12: Robustness Checks: Testing for Breaks in the Achievement Control Variables in Grade 10 ......................................................... 98
Table A.2.13: Robustness Checks: Testing for Breaks in the Control Variables in Grade 10 - Ethnicity and Gender ......................................................... 99

Table 3.1: Distribution of Students Across Years and Grades in the Longitudinal Data ......................................................... 138
Table 3.2: Special Assessment Test Design ......................................................... 138
Table 3.3: Out-of-Sample Predictive Power of the IRT Model in the Cross Sectional Data ......................................................... 139
Table 3.4: Within-Sample Predictive Power of the IRT Model in the Longitudinal Data ......................................................... 139
Table 3.5: Fraction of Students At or Above a 50% Score in the 5-Year Cohort ......................................................... 140
Table 3.6: Pair-wise Leven’s Test for Equality of Variances of the Ability Scores in the 5-Year Cohort ......................................................... 140
Table 3.7: Evolution of Student Ability Scored Across Grade Levels in the Longitudinal Data ......................................................... 141
Table 3.8: Distance of Student’s EAP Ability Score from the Grade Level Standard in the Longitudinal Data in Grades 2-5 ......................................................... 142
Table 3.9: Gains on Grade Appropriate Questions from Additional Years of Schooling in the Longitudinal Data in the 5-Year Cohort ......................................................... 143
Table 3.10: Evolution of Student Ability Scored Across Years in the Longitudinal Data ......................................................... 143
ACKNOWLEDGEMENTS

I would like to acknowledge Professor Julian R. Betts for his support as the chair of my committee. Through multiple drafts and many research meetings, his guidance has proved to be invaluable.

I would like to acknowledge my thesis committee: Prof. Julian R. Betts, Prof. Craig T. McIntosh, Prof. Karthik Muralidharan, Prof. Prashant Bharadwaj, and Prof. Amanda Datnow for their involvement in shaping of this dissertation.

I would also like to thank Prof. Julie Cullen, Prof. Gordon Dahl, Prof. Roger Gordon, Michael Kuhn, Dallas Dotter and seminar participants at UC San Diego, Department of Economics for all the valuable comments and suggestions.

Finally, I would like to extend further thanks to Dr. Karen Volz Bachofer and Andrew Zau for the technical and institutional support and to the San Diego Unified School District for the data used in the analysis and administrative support.

Chapter 2 is currently being prepared for submission for publication of the material. The dissertation author was the primary author of this material. Zieleniak, Jedrzej; Betts, Julian R. “Effects of Test Failure on Academic Achievement in the Context of Student Accountability: A Regression Discontinuity Analysis.”

Chapter 3 is currently being prepared for submission for publication of the material. The dissertation author was the primary co-author of this material. Muralidharan, Karthik; Zieleniak, Jedrzej. “Chasing the Syllabus: Measuring Learning Trajectories in Developing Countries with Longitudinal Data and Item Response Theory.”
VITA

2008 B. A. in Economics *Honors 1st Class*, Simon Fraser University, Canada

2008-2014 Graduate Teaching Assistant, University of California, San Diego

2009 M. A. in Economics, University of California, San Diego

2014 Ph. D. in Economics, University of California, San Diego
ABSTRACT OF THE DISSERTATION

Essays on Education Policy

by

Jedrzej Zieleniak

Doctor of Philosophy in Economics

University of California, San Diego, 2014

Professor Julian R. Betts, Chair
Professor Craig T. McIntosh, Co-Chair

My dissertation is a collection of essays on effects of educational reforms on student outcomes. Chapter 1 investigates the effects of the California High School Exit Exam on student achievement along the entire student distribution. I report three main findings. First, the exit exam lowers graduation rates among students in the left tail of the achievement distribution. Second, the initial, more difficult version of the exam is associated with positive effects on achievement across the student distribution, with the exclusion of the lowest performing students. Finally, the current version of the exam is associated with negative effects on student test scores and grades among the low achieving students, and with increased achievement in the middle of the distribution.
Chapter 2 analyzes the effects of initial failure on the California High School Exit Exam on student outcomes using the Regression Discontinuity framework. For students close to the passing cutoff, the initial failure does not affect student achievement while in high school. However, failing either component of the test increases the likelihood of enrollment in remedial classes the following year. An analysis of postsecondary outcomes shows that conditional on graduating from high school; initial CAHSEE failure has no effect on 2-year college enrollment or the likelihood of transfer to a 4-year university. Chapter 2 shows that the California High School Exit Exam does not pose a significant hindrance to student achievement for students close to the passing cutoff.

Combining a unique annual longitudinal dataset on learning outcomes with a vertically-scaled item-response theory model, Chapter 3 presents the first estimates of learning trajectories over time for a given cohort of students in a developing country. We provide the first out of sample validation of a vertically-scaled IRT model in a low-income setting. We document increasing inequalities in learning outcomes as students go through primary school. We document significantly increasing variance in learning over time and increasing gap between the mean learning levels and the standard specified by the syllabus over time. Our results suggest that education systems in developing countries have not effectively made the transition from serving a screening function to providing human capital to all students.
Chapter 1

Jumping Through Hoops: Student Achievement and the California High School Exit Exam

Jedrzej Yendrick Zieleniak\textsuperscript{1}

\textsuperscript{1}University of California, San Diego, Department of Economics, 9500 Gilman Drive, La Jolla, CA 92093, USA. E-mail: jzieleni@ucsd.edu. Web page: http://econweb.ucsd.edu/~jzieleni.
1.1 Abstract

Using models of academic standards in the context of high-stakes testing, I investigate the effects of the California High School Exit Exam (CAHSEE) on student achievement in the second biggest school district in California. I report three main findings. First, the exit exam lowers graduation rates among students in the left tail of the achievement distribution. Second, the initial, more difficult version of the exam is associated with positive effects on achievement across the student distribution, with the exclusion of the lowest performing students. Finally, the current version of the exam, shorter and easier than the initial one, is associated with negative effects on student test scores and grades among the low achieving students, and with increased achievement in the middle of the distribution. The results highlight the fact that test-based accountability can create “winners” and “losers”.

*JEL Classifications: I210; I240; J24*

*Keywords: Education, Exit Exams, Accountability, Learning, Human Capital*
1.2 Introduction

The No Child Left Behind Act and Race to the Top are just a few of the federal policies currently pushing for greater educational accountability with the goal of improving student outcomes. By 2012 students in 26 states were required to pass a high school exit exam in order to graduate (Reardon et al. (2010)). The tests being implemented vary in the degree of difficulty and range of material, but in all cases they have common goals: to improve student learning and to strengthen the quality of signal associated with a high school diploma. The policy advocates claim that the accountability measures improve student achievement, increase parental involvement and improve the quality of schools (Jacob (2005)). Opponents of test-based accountability suggest that these exams encourage dropouts, narrow the curriculum and put a disproportionate burden on low performers.\(^2\) The controversial nature of the test-based accountability policies in education motivates a growing literature concerning the impacts of school accountability policies on student achievement (e.g., Bishop (1998); Figlio (2006); Jacob (2005); Ou (2010)). However, there is relatively little evidence concerning how test-based accountability alters the distribution of outcomes of students exposed to the policy.\(^3\) Furthermore, the heterogeneity of effects of exit exams at different points of the test score distribution remains largely unexplored.

The literature on educational standards predicts three main effects associated with an introduction of test-based accountability (Betts and Costrell (2001); Jacob (2005)) First, the exit exam might negatively affect the poor performers through a discouragement effect.\(^4\) Second, test-based accountability might induce students

\(^2\)The discussion of the literature of the negative effects of exit exams can be found in Papay et al. (2010b)

\(^3\)Figlio (2006) finds improvements in student achievement in low performing schools, Reback (2008) finds that schools focus resources on students pivotal in determining school ratings.

\(^4\)Jacob (2001) finds that exit exams increase dropouts among low achieving students. Papay et al. (2010b) find that failing an exit exam in math increases dropouts among low-income urban students in Massachusetts.
to increase effort leading to higher levels of learning through a motivational effect.\textsuperscript{5} Third, the introduction of an exit exam can induce a shift in effort towards meeting the requirements and away from other tasks.\textsuperscript{6} Teaching and learning to the test, in the most extreme case, could decrease the overall learning levels illuminating the potential distraction effects of exit exams.\textsuperscript{7} The three effects are predicted to operate in different parts of the student achievement distribution suggesting that an introduction of an exit exam might affect students at different level of achievement in very different ways. This illuminates the fact that test-based accountability might have clear “winners” and “losers” and therefore can be expected to be met with opposition.

In this paper I use a panel of student achievement data from San Diego Unified School District to study the impact of the California High School Exit Exam (CAHSEE) on academic outcomes. I analyze heterogeneity on two key dimensions. First, I investigate how the impacts of the exam vary, based on student’s middle school achievement and position in the achievement distribution. Second, I take advantage of student selection into math courses with varying overlap with the exit exam in order to investigate the heterogeneity of effects associated with the content overlap between the courses and the exit exam. Last, I take advantage of the introduction of the CAHSEE and the subsequent changes to the exam in order to investigate the effects of variation in the exam difficulty on student achievement.

My theoretical framework introduces an exit exam in a context of a production function where students multitask. I distinguish between an introduction of an exit exam with low overlap with the grade-specific material and an exam with high

\textsuperscript{5}Hanushek and Raymond (2004) find positive effects of accountability on student performance on the National Assessment of Educational Progress across the entire achievement distribution. Jacob (2005) finds increases in achievement following an introduction of test-based accountability in Chicago Public Schools.

\textsuperscript{6}Agents are predicted to focus on directly measurable outcomes, when faced with incentive schemes based on objective criteria (Holmstrom and Milgrom (1991)).

\textsuperscript{7}Jacob (2005) finds that gains in Chicago Public Schools were driven by increases in test-specific skills.
overlap. I provide theoretical predictions of the differences between the two cases and consider distributional effects of the introduction of test-based student accountability (based on Lazear (2006)). I focus on two main questions. 1) What are the impacts of the exit exam regime on dropouts, graduation rates and test performance across the different parts of the student achievement distribution? 2) What are the impacts of the exam regime on different domains of learning? Are those domains not covered on the exam negatively affected?

I take advantage of the pre-existing school accountability system, the California Student Testing and Reporting (STAR) system, in order to construct a student academic achievement index. I use the index to investigate the differential effects of the exit exam regime in a triple difference framework. The first difference is the achievement before and after the introduction of the exit exam. The second difference is between students required to take the exam for the first time (grade 10) and students not yet exposed to the exam (middle-schoolers in grades 6 through 8). The third difference is the comparison between students at different points in the achievement distribution. I consider the heterogeneity of effects along the student achievement distribution showing the difference in effects between the two exit exam regimes introduced in the state and how they relate to learning.

I report two main sets of findings in this paper. First the exit exam increased dropout rates among students in the first and third decile of the achievement distribution by 1 percentage point, when the exam was a graduation requirement. That is a 25% increase from the pre-policy period. Furthermore, the exit exam requirement is associated with an average drop in probability of on-time graduation of 35.7 percentage points among the bottom 30 percent of students. While the exit exam
delays graduation for the bottom three deciles of the achievement distribution, only the lowest performing students’ graduation rates are affected in the longer term. I find that the exit exam requirement reduces the graduation rates of the bottom ten percent of students by 25 percentage points.\(^\text{10}\)

Second, the overlap of material between the exit exam and course work has a positive effect on student achievement on standardized tests. For students enrolled in Algebra 1, the exit exam is associated with an increase in math test scores of between 0.16 and 0.258 of a standard deviation, with the gains being even larger for students in Algebra 2. In contrast, for students enrolled in Geometry, the exit exams is associated with decreases in tests scores of 0.125 of a standard deviation. Similar patterns can be observed in terms of students’ grades in math courses, however only the negative effects associated with Geometry are statistically significant. For students enrolled in Geometry in grade 10, the exam lowers math grades by 0.281, roughly 10% of the mean grade-point average. The effects associated with Algebra 1 and Algebra 2 are not statistically significant.

A key advantage of the analysis presented in this paper, in comparison to most recent studies investigating the effects of exit exams, is the fact that the empirical approach allows to estimate the effects of the exams along the entire achievement distribution. In contrast, the most recent studies investigating test-based accountability employ the regression discontinuity framework (e.g. Reardon et al. (2010), Ou (2010), Papay et al. (2010b)). The RD design allows for strong identification, but is only relevant to a small fraction of students in the close neighborhood of the passing cutoff. My empirical approach allows for estimation of exam effects for students at any point in the distribution. The estimation strategy relies on the parallel trends assumption between students in middle school and in high school in the pre-exam regime. I test

\(^{10}\)Graduation is defined as the probability of obtaining a high school diploma within five years of enrolling in grade 9 for the first time.
the parallel trends assumptions using pre-exam data and cannot reject parallel trends at 5% level for any of the achievement outcomes. Furthermore, I assume that the changes in achievement among students in middle school are orthogonal to the introduction of the exit exam.

The analysis presented in this paper provides evidence on the effects of test-based accountability on student achievement. However, there are potential positive effects of the exam introduction which are not quantified here. The exam could improve the strength of the signal associated with the high school diploma. The data available in the analysis does not allow for estimation of potential employment and wage effects. Thus, while this paper provides an analysis of the effects of test-based accountability on academic performance, it is by no means exhaustive in terms of consideration of all effects associated with the exam.

1.3 Institutional Context

1.3.1 California High School Exit Exam

The initiative to introduce a comprehensive program of accountability started in California with the passage of the Public School Accountability Act of 1999. The legislation led to the creation of subject content standards, and a state-wide system of student high-stakes testing. CAHSEE is a key component of the accountability system. The exam is the only part of the system with consequences for students. Failure to pass the exam by the end of grade twelve prevents students from obtaining a high school diploma. The exam consists of two independent tests: mathematics and English Language Arts (ELA). The ELA component encompasses material relevant to grades six through ten, while the math test is designed to test material covered in
grades six through eight. All grade ten students are required to take the CAHSEE. Students have up to six attempts at taking the test. The important feature of the CAHSEE is that once a student passes a given component of the exam, either math or ELA, only the component failed in the past has to be retaken. The tests are designed with a single passing cutoff of 350 points for either of the tests and there is no additional recognition for scoring exceptionally well.\textsuperscript{11}

The test requirement was introduced in 2001, with the expectation that starting with the students of the graduating class of 2004, all students in California would be required to pass the CAHSEE in order to obtain a high school diploma. The initial version of the test was longer and covered a wider set of topics than the current version of the exam. The initial exam was administered in years 2001-2002 through 2002-2003. In that period the graduating classes of 2004 and 2005 took the exam, considering it a graduation requirement. I will refer to the early exam as CAHSEE 1. The material to be tested on CAHSEE 1 included a wider range of topics than the subsequent version introduced in 2004. In the summer of 2003 The State Board of Education announced that the CAHSEE requirement would be suspended and the test would be revised and re-introduced starting with the graduating class of 2006. The decision to temporarily withdraw the requirement from the accountability system was dictated by a large fraction of students entering grade twelve in the school year 2003-2004 who were still to satisfy both components of the test.

The exam was reintroduced as a graduation requirement starting with the graduating class of 2006. Acting based on the advice from the State Board of Education, the California Department of Education reduced the test length and testing time and modified the exam’s content. ELA testing was reduced from two to one day,\textsuperscript{11} The literature suggests that the incentive effects in a minimum-competency test are concentrated among students who must repeat the test as the initial failure might stimulate more learning Bishop (1998)
mainly by removal of one of the two essays required on the exam. Additionally the number of multiple choice questions asked on the exam was reduced from 82 to 72. The Education Testing Service reported that the changes to the ELA test maintained equal weighting of content standards between reading and writing. The length of the math component of the exam remained constant at 80 multiple choice questions, but the difficulty of the material tested was refined (Service (2005)). The changes in the exam raised concerns that the new version of the exam was pitched at too low of an achievement level and that it would do nothing to increase effort and learning of the low-performing students (Zau and Betts (2008)). The new version of the exam was introduced in the 2003-2004 school year and the test continues to be a requirement for students in California. The current version of the exam will be referred to as CAHSEE 2.

The policy changes associated with the introduction of CAHSEE imply that students in grade 10 in different school years faced different incentives. Students in grade 10 prior to the 2001-2002 school year did not face an exit exam as a graduation requirement. In contrast, the 10th graders in 2001-2002 and 2002-2003 were faced with the CAHSEE 1 requirement, while the 10th graders starting with year 2003-2004 were required to pass CAHSEE 2 in order to graduate. Table 1.1 summarizes which grade levels in what years were required to pass CAHSEE and which version of the exam was required.

TABLE 1.1 HERE

CAHSEE is not the only graduation requirement faced by students in California. All students are required to successfully complete a set of 13 year-long courses
including English, mathematics and science.\textsuperscript{12} School districts impose additional graduation requirements such as minimum grade-point average.\textsuperscript{13} The exit exam should not be viewed as the only potential barrier to graduation, but rather as an additional hurdle introduced to monitor and assure the graduating students meet the specified standard. An important feature of the pre-existing requirements is the fact that students across all grades in high school are directly accountable for maintaining an adequate grade-point average and completing the required courses, but only the students at specified grade levels are exposed to the exit exam requirement.

\subsection*{1.3.2 Student Testing and Reporting (STAR)}

Separate from the exit exam, the school accountability testing was introduced in California in 1998 as part of the Student Testing And Reporting System. The tests introduced as part of STAR are administered to students in grades 2 through 11. The results from standardized tests are used to calculate the Academic Performance Index (API), used in the state accountability system. The standardized tests carry no consequences for the students, making them low-stakes. However, test results are used in the school accountability system making the same tests high-stakes for teachers and administrators.

The tests are administered and regulated under the California Standardized Testing and Reporting system and were introduced in the 1997-1998 school year. Initially the state used the Stanford Version 9 (SAT9) test, but it was replaced in the 2001-2002 school year by the California Standards Test (CST). The notable difference between SAT9 and CST is the fact that SAT9 was grade specific for all subjects covered. In contrast, the math CST is course specific. This implies that a students

\textsuperscript{12}California Education Code (EC) 51225.3
\textsuperscript{13}The San Diego Unified School District requires its students to attend eight semesters for six periods daily and a cumulative grade-point average of 2.0 or higher
in grades 8 through 11 have a choice of which math class to enroll in and therefore which CST they will have to take. The change in the standardized test was implemented with a single year of overlap, where all students in grades 2 through 11 in the 2001-2002 school year were required to take both the SAT9 and the CST. The relationship between the test scores on both tests by grade level in math and ELA are presented in Figures 1.1 and 1.2. The scores on SAT9 and CST in both English and math reveal the same pattern of a strong positive correlation. The change in the testing structure in math can also be observed as the data reveals a larger degree of dispersion starting in grade eight, relative to the correlations in earlier grades. The fact that both tests were administered in the 2001-2002 school year will be used as the test scores will be normalized relative to the population in 2002, the students who were exposed to both of the standardized tests.

FIGURE 1.1 HERE
FIGURE 1.2 HERE

1.4 Data

The data used in the empirical analysis were obtained from the San Diego Unified School District (SDUSD). The data is an unbalanced panel of student level observations for grades six through twelve for the 1997-1998 to 2009-2010 school years. The analysis of heterogeneity of effects by initial student achievement restricts the sample to be used in the analysis to the students who are observed in grade eight. The sample restrictions imply that a subsample of 408,561 student-year observations will be used in the analysis with further restrictions in sample size based on the availability of a full set of controls and the outcomes of interest.

The outcomes capturing achievement in high school are the student scores on
standardized tests administered in California under the Standardized Testing and Reporting (STAR) in English Language Arts and in mathematics, as well as the student grade-point average in math and English in grades 6 through 12. Student low-stakes test scores are standardized at every grade level to the mean and standard deviation of the student population in SDUSD in the 2001-2002 school year when both SAT9 and CST were administered. Thus, the empirical analysis of the effects of exit exams on test scores estimates effects relative to the 2001-2002 student population. I further analyze the effects of exit exams on the probability of dropping out, on-time graduation and the likelihood that a student graduates within five years of enrolling in grade 9 for the first time.

The set of controls used in the analysis contains information on the courses taken in English and math as well as the lagged values of grades obtained. I observe student’s grade-point average, gender, age, ethnicity, special education status, English Learner status, the percentage of the school year the student was absent, and parental education level. School characteristics used in the analysis include the school ethnic composition as well as the fraction of students on free or reduced-price lunch. For any control variables also used as an outcome of interest the set of controls contains lagged values of the outcome, thus the empirical models can be viewed as models of gains rather than levels.

The administrative data can be divided by the type of an exit exam regime faced by the students. The first category is the students enrolled in SDUSD when CAHSEE was not a requirement, referred to as “No CAHSEE”. Second is the students enrolled in middle or high school when CAHSEE 1 was required, with the requirement being lifted in the summer of 2003. The third group is the students enrolled in middle or high school when CAHSEE 2 was a requirement. CAHSEE 2 encompasses the students who ultimately were required to pass the exit exam in order to
graduate. Table 1.2 summarizes the group characteristics for students by regime type.

TABLE 1.2 HERE

The data shows that the students across all regimes are similar in terms of grade-point average, math and English grades. Student scores in math and ELA exhibit increases as the educational system progressed from no exit exam, through CAHSEE 1 to CAHSEE 2. The increases in scores reflect the fact that the educational system improved over time and the scores on standardized tests, relative to the student population in school in 2001-2002 should be improving. The data also reveals that the fraction of English Learners in the district decreased over time. Table 1.2 shows that in general performance in the district is improving as measured by test scores.

The empirical strategy employed in this paper relies on the assumption of parallel trends between students in middle school and in grade 10 before the introduction of the exit exam. I test the parallel trends assumption for achievement in school and present the estimation results in Table 1.3. The parallel trends assumption holds for all outcomes presented, as the coefficients on the interaction of a linear time trend and a grade 10 dummy variable are statistically not significant across the four outcomes considered.

TABLE 1.3 HERE

1.5  Empirical Framework

The effects of exit exams on learning are one of the motivations for the introduction of such exams. However, the introduction of an exit exam would also affect
students likelihood of graduation and dropping out. The theoretical characterization of the effects of an increase in standards and the introduction of an exit exam is presented in Appendix A. Hybrid models of human capital/signaling predict that if the exit exam increases the academic standards then the student population can be divided into four distinct groups. 1) Students at the bottom of the achievement distribution, who did not meet the pre-exam graduation requirements. 2) Students who in the pre-policy period would have obtained a diploma but fail to do so under the new regime. 3) Students still meeting the requirements, who might experience the motivational effects of the exam. 4) Top achievers for whom the introduction of the exit exam should provide no incentives to change behavior.

The theoretical predictions related to learning and graduation rates will be tested against the student level data. Theory suggest that the exam should have a negative effect on graduation and persistence in school for students towards the bottom of the distribution. For students meeting the standards the introduction of the exit exam should be associated with increased achievement due to the motivational effects and the uncertainty of what will be tested. Furthermore, the achievement gains should be larger in relation to CAHSEE 1 as compared to CAHSEE 2, and in both regimes among students enrolled in Algebra courses where the overlap of material with the exam is the biggest, as compared to those in Geometry and other courses.

1.5.1 Student Achievement While in High School

The effects of the exit exam on performance on standardized tests might depend on the range of the exit exam (the knowledge domain that is fair game to be tested) as well as the overlap of the exam with the grade level material. Furthermore, the effects of the exit exam are predicted to depend on student’s position in the

\(^{14}\)Their achievement should be lower than achievement of comparable students pre-CAHSEE due to the discouragement effect.
achievement distribution. I use student characteristics in grade 8 (the pre-treatment grade level) in order to estimate the achievement distribution among the students in relation to the exit exam. For the student population exposed to the exit exam regime I model the lower of the ELA or Math scores obtained by student i on the CAHSEE in grade 10, min(CAHSEE)\textsubscript{g\textsubscript{10}}\textsuperscript{g\textsubscript{10}} \textup{ist}, using student’s grade 8 characteristics:

\[
\min(CAHSEE)_{\text{g10}}^{\text{g10}}\text{ist} = \beta X_{\text{is}(t-2)} + \epsilon_{\text{ist}}\tag{1.1}
\]

where \(X_{\text{is}(t-2)}\) is a vector of student characteristics in grade 8, including low-stakes math and ELA scores measured as Z-scores normalized to the mean and standard deviation of grade specific performance in the 2001-2002 school year, absences, English learner status, gender and ethnicity and \(\epsilon_{\text{ist}}\) is a stochastic error term. I use the model to predict the lower of the CAHSEE scores for all students in my sample whom I observe in grade 8, including those who were never required to pass the exit exam. I use the predicted scores to classify students into deciles of performance of an achievement index (\(AI_i = \hat{\beta} X_{\text{is}(t-2)}\)). The deciles will be defined according to the distribution of predicted scores among students in grade 10 under the CAHSEE 2 regime. In this sense, the deciles will not be relative to one’s peers but rather relative to the distribution of achievement among 10\textsuperscript{th} graders under the current exit exam regime, so that none of the students in the study will be switching classifications across grade levels.

I assume that the exam affects student performance directly only in the grade levels in which students are required to take the exam. Furthermore, I allow for the two versions of the CAHSEE to have different effects on student achievement. In order to address the potential concerns about students’ forward looking behavior I control for strategic behavior in grade 9, one grade level prior to when the student has to take the exit exam. Theoretical predictions motivate the focus on the analysis of heterogeneity of effects along the student academic achievement dimension. I address the achievement heterogeneity in a triple difference framework.

I estimate a triple difference model of achievement among students in grades 6 through 11 (with grade 12 included in the analysis of grades and GPA), where \(Y_{\text{idgst}}\)
is achievement on a standardized tests, or grades in math or ELA, for student i in decile d of the Achievement Index in grade g at school s in year t:

\[
Y_{idgst} = \beta_0 + \sum_{j=1}^{2} \sum_{b=1}^{10} \left[ \beta_{jb} \text{Ex}(j)_{-10, D(b)_{idgt}} + \tau_{jb} \text{Ex}(j)_{-10, D(b)_{idt}} \right] + \sum_{b=1}^{10} \left[ \zeta_b \text{Grd}10_{-10, D(b)_{idgt}} \right] + \sum_{b=1}^{10} \left[ \xi_b \text{Grd}10_{-10, D(b)_{idg}} \right] + \gamma X_{idgst} + \theta Z_{ist} + \delta_s + \delta_t + \epsilon_{idgst}
\]  

(1.2)

Ex(j)_{-10, D(b)_{idgt}} is a binary variable equal to one if student i was in grade 10, in decile b when Exam-CAHSEE 1 (Ex(1)=1) or CHASEE 2 (Ex(2)=1) was a requirement, e.g. Ex(1)_{-10, D(b)_{idgt}} is equal to one if student i was in grade 10 in the year 2001-2002 or 2002-2003 and student was classified in decile 3. The coefficients \( \beta_{jd} \) are the differential effects of being required to take the CAHSEE in grade 10 under one of the two regimes for students in different deciles of the achievement index. Those are the triple difference coefficients of the differential effects of the CAHSEE regime that are of interest for this analysis.

Ex(j)_{-10, D(b)_{idt}} is a binary variable equal to one if student was in decile b and enrolled in school in years 2002 or 2003 if j=1 (under the CAHSEE 1 regime) or in years 2004 or later if j=2 (under the CAHSEE 2 exit exam regime). Grd10_{-10, D(b)_{idt}} is a binary variable equal to one if student i was in grade 10 and classified as being in decile d of the academic achievement index, based on student achievement in grade 8. D(b)_{idt} is a binary variable equal to one if student i was classified as being in decile d of the achievement index. Ex(j)_{-11,igt} is a binary variable equal to one if student i is in grade 11 in year 2003 if j=1 (CAHSEE 1 regime) or in year 2005 or later if j=2 (CAHSEE 2 required). The reason why the variable is not equal to one for students in grade 11 in year 2004 is because the CAHSEE 1 requirement was lifted in the summer of 2003 and students in grade 11 in 2004 were not required to take the CAHSEE.
Ex(j)Grd9igt is a binary variable equal to one if student i is in grade 9 in year 2002 or 2003 if j=1 (CAHSEE 1 regime) or in year 2004 or later if j=2 (CAHSEE 2 required). $\beta_{4j}$ captures the potential forward looking behavior of ninth graders in relation to the exit exam. CAHSEE(j)$_t$ is a binary variable equal to one if year equals to 2002 or 2003 for j=1, and 2004 and later for j=2. $X_{igst}$ is a vector of student characteristics, $Z_{st}$ is a vector of school characteristics, $\delta_g$ is a grade fixed effect, $\delta_s$ is a school fixed effect, $\delta_t$ is a year fixed effect and $\varepsilon_{igst}$ is a stochastic error term.

The coefficients $\beta_{jb}$ should capture the differential effects of the exit exam in grade 10 by ability level and by the exit exam regime. Similar empirical framework will be used to explore the issue of the overlap of the material with standardized tests. Theory predicts that students in courses more aligned with the CAHSEE have stronger incentives to master the material testing the course content than students in courses with little overlap with the exam. I investigate this hypothesis taking advantage of the courses taken by students in mathematics. I estimate a triple difference model of differential effects of the exit exam based on the math course taken in grade 10 parallel to the model formalized in Equation 1.2 with the decile dimension replaced by controls for the math course taken by the student in grade 10.

### 1.5.2 Effects of the CAHSEE on Graduation Rates

Estimation of the effects of the CAHSEE regime on graduation and dropout rates further takes advantage of the academic achievement index defined above. I estimate a difference-in-difference model where the first difference is the difference in graduation rates before and after the introduction of the CAHSEE and the second difference is between students in different deciles of the predicted achievement distribution. I model $Y_{idst}$, the probability that student i in decile d at school s in year t graduates with a diploma, graduates from high school on time or drops out. The
administrative data does not provide information whether every student missing in the data is a dropout as some students move from the district. This implies that dropouts as defined in the data are not mechanically connected to graduation as all students who do not graduate. Thus, dropouts and graduation rates will be modeled separately using the following probability model:

\[
Pr(Y_{idt} = 1) = \Phi \left( \beta_0 + \sum_{j=1}^{2} \sum_{b=1}^{10} \beta_{jb} \text{Ex}(j)D(b)_{idt}^{Al} + \sum_{j=1}^{2} \alpha_j \text{Ex}(j)_t + \sum_{b=1}^{10} \tau_b D(b)_{id}^{Al} \\
+ \gamma X_{idst} + \theta Z_{st} + \delta_t + \delta_s + \nu_{idst} \right)
\]  

(1.3)

where \( \text{Ex}(j)D(b)_{idt}^{Al} \) is a binary variable equal to one if student \( i \) in decile \( d \) was in grade 12 in year 2004 or 2005 if \( j = 1 \), and in 2006 or later if \( j = 2 \). The differential effects of CAHSEE 1 by decile are captured by the \( \beta_{1b} \) coefficients. CAHSEE 1 was ultimately not a graduation requirement, but until the summer of 2003 the students in the graduating classes of 2004 and 2005 were convinced that it would be. The estimated effects associated with CAHSEE 1 therefore will not capture any direct barrier to graduation due to failing the exit exam, but should rather be thought of as the total effects of all the factors that might have been affected by CAHSEE 1 and which might have significance towards graduation. The set of the \( \beta_{2b} \) coefficients captures the differential effects associated with CAHSEE 2 by decile of the achievement index. CAHSEE 2 remained a graduation requirement, therefore the estimated effects associated with CAHSEE 2 should be larger in magnitude than the effects associated with CAHSEE 1.

\( \text{Ex}(j)_t \) is a binary variable equal to one if year equals 2004 or 2006 if \( j = 1 \), and if year greater or equal to 2006 for \( j = 2 \), \( D(b)_{id}^{Al} \) is a binary variable is student \( i \) is classified in decile \( d \). Vector \( X_{idst} \) is a set of student characteristics, vector \( Z_{st} \) is the vector of school time-varying characteristics to control for changes in the schools
themselves. Furthermore, I introduce $\delta_t$, the year fixed effects, and $\delta_s$, school fixed effects in order to control for all sharp differences in graduation rates across years and schools.

## 1.6 Results

### 1.6.1 Student Achievement on Standardized Tests

Figure 1.3 presents the comparison of test scores and grades of students in grades eight and ten by school year. The two vertical lines signify the introduction of CAHSEE 1 and CAHSEE 2 respectively. All four panels show that the trends between grade 10 and grade 8 performance co-moved really closely in the pre-policy period. With the introduction of the exit exam, the pattern of performance between grades 8 and 10 diverges, with grade 10 performance exhibiting significantly lower levels of growth. The panel of mathematics scores reveals the largest deviation of grade 10 scores from the pre-policy trends. The large deviation is partially associated with the fact that SAT9 was a grade-specific test, while the CST is course specific. Thus, students enrolled in classes more difficult than the expected grade level material might perform worse on standardized tests than comparable students taking the SAT9 tests. The ELA tests remained grade-level specific, between the SAT9 and the CST. The ELA scores provide a more accurate picture of the potential effects of the exit exam introduction on student test scores, as presented in the top right panel of Figure 1.3.

FIGURE 1.3 HERE

The pattern in math test scores is also visible in math grades - grade 10 per-
formance co-moves with grade 8, but with the introduction of the exit exam, the achievement across the two grade levels diverges, with continuing growth in grade 8 and much slower improvements in grade 10. Here, the content being evaluated is grade specific, therefore the divergence after the introduction of the CAHSEE cannot be attributed to the change in the low-stakes testing environment. The potential worry in relation to grading under the CAHSEE regime would be that teachers might be more lenient towards lower performing students. I cannot rule out the change in teacher behavior and estimated effects of the CAHSEE should be viewed as the overall effect of the exit exam.

The model formalized in equation (1.2) is used in order to estimate the differential effects of CAHSEE 1 and CAHSEE 2 on student achievement on the low-stakes standardized tests. The estimation results are presented in Table 1.4. Columns 1 and 2 refer to the same model of ELA scores, while columns 3 and 4 to the model of math scores. Furthermore, the coefficients, along with their confidence intervals are also presented in Figure 1.4

The regression analysis suggests that the effects of CAHSEE 1 and CAHSEE 2 on standardized tests are vastly different from one another. Overall, CAHSEE 1 seems to be associated with positive effects on ELA and math scores. At the bottom of the achievement distribution, in decile 1 (based on student achievement in grade 8) CAHSEE 1 is associated with no gains in math, but a 0.095 of a standard deviation increase in ELA scores as compared to the difference in grade 8 scores. The gains can further be detected across deciles 2 through 10 in both math and ELA. CAHSEE 1 is associated with an ELA score increase of 0.272 of a standard deviation in decile 10
and a corresponding increase of 0.478 of a standard deviation in math. The regression analysis estimates statistically significant increases in math for deciles 6 through 9. As mentioned before, the low-stakes testing in math changed considerably in 2002. The system moved towards course specific testing, therefore students might be taking low-stakes tests which are more difficult, or easier than before. In order to address that change I control for the student math class taken in the new low-stakes testing regime. The set of controls might not capture the change in the low-stakes testing regime fully, however, it is dictated by the institutional change and data limitations. The estimated coefficients associated with math test scores should therefore be taken with caution and viewed as a joint effect of the introduction of the exit exam, and the change in the low-stakes testing environment.

The estimated effects associated with CAHSEE 2 are more in line with the theoretical predictions than the effects of CAHSEE 1. The effects on ELA and math scores (columns (2) and (4) respectively) indicate that the CAHSEE 2 regime is associated with a decrease in performance among students in grade 10 and in the bottom 4 deciles for ELA and bottom 5 deciles for math, as compared to students at the bottom of the achievement distribution in grade 8. Students at the bottom of the distribution experience a 0.204 of a standard deviation score decrease in ELA and a 0.337 of a standard deviation decrease in math scores. The negative coefficients are in line with the theoretical predictions of a discouragement effect being localized to the students at the bottom of the achievement distribution. Furthermore, the exit exam should have no effect at the right tail of the achievement distribution. In both ELA and math, the effects associated with CAHSEE 2 in the top deciles are statistically insignificant. CAHSEE 2 is associated with no effect of the regime on ELA test scores among students in deciles 5 through 10. Similarly, the effects of CAHSEE 2 on math scores for deciles 6 through 9 are not statistically significant. The results presented
in Table 1.4 are also graphically summarized in Figure 1.4.

FIGURE 1.4 HERE

A significant concern, when analyzing low-stakes test scores is that with the introduction of an additional requirement, students might shift effort away from standardized testing, in order to focus on the tasks required for graduation. The only part of the achievement distribution which could be indicative of a distraction effect, are the students in deciles 3 through 5 under the CAHSEE 2 regime. Those students would be expected to meet the graduation standards, including the exit exam, therefore the negative effects associated with CAHSEE 2 should not be connected to a discouragement effect. For these students it is possible that the increased effort would all be directed towards tasks that matter for graduation, and some effort previously directed towards standardized testing might have been diverted away as well.

The shift of effort away from low-stakes tests is a concern. However, we should not observe the same type of a shift away from performance in courses taken. Maintaining an adequate cumulative GPA is one of the graduation requirements therefore the incentives to do well in classes remain in place with the introduction of the exit exam. The regression results estimating the effects of exit exam on grades are presented in Table 1.5, and summarized in Figure 1.5.

TABLE 1.5 HERE

The pattern of effects associated with CAHSEE 1 in test scores is similar to the pattern observable in GPA. In both cases there is no effect of the exam among the students in the bottom decile in grade 10. For the remainder of the students the
exam is associated with positive effects on grade-point average in grade 10 ranging between 0.144 for decile 2 in grades in English, to a 0.397 increase in math grades in decile 10. In contrast, CAHSEE 2 is associated with negative effects at the bottom of the achievement distribution among students in grade 10. The exam is associated with a -0.136, a 5.6% decrease in English, and -0.253, a 12.4% decrease in math GPA.

The lack of significant effects in deciles 2 through 5 under the CAHSEE 2 regime is reassuring - the students for whom the introduction of CAHSEE 2 seems to be associated with decreases in test scores do not experience decreases in grades in math and English. This suggests that the shift of effort away from testing is a possible explanation for the effects estimated on test scores. CAHSEE 2 is further associated with increases in math and English GPA among students at the top of the achievement distribution. The effects in math are concentrated in the top two deciles, where the exam is associated with an increase in gpa of 0.171 for decile 9 and 0.202 among students in the top decile. In English the positive effects associated with the exam span the top four deciles, with estimates ranging from 0.146 for students in grade 10 classified in decile 7, to 0.274 for students in the top decile.

Figure 1.5 HERE

Under both exit exam regimes students at the top of the distribution exhibit increases in math and English grades. The finding is not consistent with the motivational predictions, as the top performers should not be changing their behavior. The increases in grades might be associated with the uncertainty of what material might appear on the exam. It is possible that the low cost learners would learn significantly more than just the material which ends up appearing on the exit exam. Alternatively, top performers typically plan to attend college. Thus, with the introduction of an
exit exam, they might increase their achievement in courses, in order to distinguish themselves further from average students who pass the exam on their first attempt.

### 1.6.2 Exit Exam and Graduation

Theoretical predictions regarding student achievement further extend to dropout and graduation rates. The students at the bottom of the achievement distribution might be discouraged from exerting effort, lowering their achievement. The same students should also be more likely to drop out of school or fail to graduate from high school, with the introduction of an exit exam. In contrast, the students in the middle of the initial achievement distribution, who continue meeting the graduation requirements might potentially experience the encouragement effects associated with the introduction of the exam. With increased motivation, the dropout rates might decrease and the likelihood of graduation might rise. Finally, the top performers should not be affected by the introduction of the exit exam, as for them the new requirement poses no challenge.

It is worth noting that the effects of the CAHSEE 1 regime on dropout rates and graduation should be viewed differently than the results associated with the CAHSEE 2 regime. The initial exit exam was removed from the graduation requirements in the summer of 2003 so the estimated effects of CAHSEE 1 on dropouts and graduation rates should be viewed as the total effects of the exam requirement and its subsequent removal.

I analyze the effects of the exit exams on dropout rates first. The smoothed raw data plots of dropout rates by exam regime and decile of the academic achievement are presented in Figure 1.6. The three regimes differ very little in the top six deciles (deciles 4-10). However, for the students in the bottom four deciles dropout rates differ between the three environments. The pre-exam dropout rates are at ap-
approximately 10% in the bottom decile of the achievement index and steadily decline as we move towards higher deciles. CAHSEE 1 is characterized by dropout rates in the bottom decile comparable to the pre-exam period, with a sharp decline in dropouts by decile 2. The lower levels of dropouts under CAHSEE 1 might seem puzzling. However, the exam was accompanied by a set of remedial policies aimed at helping the students who struggled with the new requirement. It is possible that the remedial help and the increased attention from the teachers and schools towards struggling students prevented a significant number of those students from dropping out. Furthermore, CAHSEE 1 did not end up being a binding requirement, which also might have affected student decision whether or not to drop out.

FIGURE 1.6 HERE

Dropout rates under the CAHSEE 2 regime follow the general pattern observed in the other regimes. However, dropout rates appear approximately 3 percentage points higher than in the pre-exam period at the bottom of the achievement index. The local polynomial smoothed dropout rates do not control for other observable characteristics, therefore the observations made based on raw data might not translate into significant increases in dropouts in a regression framework.

In order to analyze the potential impacts of the exit exam regime on dropout rates, I use model 1.3, as described above. I use a full set of controls including student and school characteristics, measures of student past performance, as well as year and school fixed effects. The regression results are presented in Table A.2.5. The coefficients in Table A.2.5 are the difference-in-difference estimates, where CAHSEE 1 and CAHSEE 2 regimes are interacted with the deciles of the achievement distribution. Each column corresponds to the decile of interest and all the coefficients are estimated
within a single regression framework, using the richest possible model.

TABLE A.2.5 HERE

The regression results suggest that the CAHSEE 1 regime was not associated with increases in dropout rates. It has to be noted that the initial exam regime is associated with a decrease in dropout rates of 1.1 percentage points from the pre-exam level for students in decile 6, a finding suggestive of motivational effects. The possible explanation for the estimated zero effect at the bottom of the distribution is the focus of schools and teachers on the students at risk. This increased attention could have been sufficient to prevent students at risk of failing the new requirement from dropping out. Ultimately, the CAHSEE 1 requirement was lifted in the summer of 2003. The students who continued to fail the exam would be likely to drop out in grades 11 or 12. However, with the removal of the exit exam, the incentives to leave school were significantly reduced for students who completed grades 10 or 11.

The CAHSEE 2 requirement, as opposed to the CAHSEE 1, remained in place. For some students, the exam could have been the main barrier to graduation. The regression results suggest that CAHSEE 2 is associated with increases in dropout rates for students in deciles 1 and 3. The exit exam is associated with a 1 percentage point increase in the likelihood of dropouts among students in the first or third decile of the achievement distribution. While the coefficient associated with decile 2 (row 2, column 2) is not statistically significant, the sign of the coefficient is consistent with an increase in dropout rates. Furthermore, the exam is associated with lower dropout rates among the top performers. The results associated with CAHSEE 2 among the top performers do not align with the theoretical predictions. The estimates suggest that with the introduction of the exam, the best students are less likely to drop out. It
is possible that the results are affected by the very conservative definition of dropouts used in the analysis.

The analysis of dropout rates suggests very small effects of the exit exam regimes. If the exit exam was a barrier to graduation for some of the students at the bottom of the achievement distribution, the students who did not drop out should experience lower graduation rates or at least delays in graduating from high school. The legislation introduced in response to the exit exam provided schools with additional funding allowing students to enroll in school past grade 12 in order to complete graduation requirements. I now use the model formalized in equation 1.3 in order to analyze the effects of the exit exam regimes on the probability of graduating on-time.

Figure 1.7 presents the relationship between deciles of the achievement index and the probability of on-time graduation by the exam regime type. The on-time graduation rates among students classified in deciles two through four are higher during CAHSEE 1 regime, as compared to the pre-exam period. This is consistent with student expectations about the exam as difficult and an increases in academic effort. This interpretation is consistent with the previously estimated positive effects on standardized tests and grades. Additionally, those students could have been experiencing increased attention from teachers, being identified as potentially at risk of struggling with the exam. The estimated increases in performance, coupled with a lift of the exam requirement could have resulted in increases in on-time graduation— an unintended consequence of backing out of the initial policy.

FIGURE 1.7 HERE

The pattern associated with CAHSEE 2 is significantly different than that for CAHSEE 1. Students in the bottom two deciles, required to pass CAHSEE 2,
have significantly lower probabilities of on-time graduation than similar students in the pre-exam period. The pattern is consistent with the exam being a challenge for students at the bottom of the achievement distribution. The exam appears to delay students’ ability to meet the graduation requirements, resulting in decreases in the on-time graduation rates. The corresponding regression results are presented in Table 1.7.

The effects associated with CAHSEE 1 appear to be in line with a strong motivational effect across all but the bottom of the achievement distribution. The results are consistent with the findings associated with standardized tests and grades. The initial exam regime is associated with a 1.9 to 2.9 percentage points increases in the likelihood of on time graduation. The regression estimates suggest that the early version of the exam motivated students across the entire achievement distribution. A larger fraction of the student population managed to meet the graduation requirements on-schedule. The estimated response pattern suggests that students might have been uncertain about the difficulty of the exam and might have exerted more effort than they would have in absence of the exam.

TABLE 1.7 HERE

The effects associated with the CAHSEE 2 regime follow the theoretical predictions more closely. The later exam regime appears to be associated with strong negative effects on the probability of graduating on time in the bottom three deciles. Students in the three bottom deciles experience an average of 35.7 percentage point decrease in the probability of on-time graduation. The probability of on-time graduation among the top achievers appears not to be affected by the introduction of CAHSEE 2. The findings suggest that an easier exit exam decreased on-time gradu-
ation rates while the more difficult exam is associated with positive effects. The fact that CAHSEE 1 was rescinded and did not become a graduation requirement implies that students working towards graduation did not have to actually pass the exam. In contrast those required to take CAHSEE 2 were required to pass it.

While the CAHSEE 2 appears to be associated with delays in meeting the graduation requirements within the expected time frame, ex-ante it is not clear whether the delays result in significant decreases in the overall probability of students graduating within five years of enrolling in grade 9 for the first time. Figure 1.8 presents plots of graduation rates as a function of the deciles of the achievement distribution\textsuperscript{15}. The pattern associated with graduation rates is similar to the findings associated with the probability of on-time graduation, with the differences between the regimes being somewhat smaller. The likelihood of obtaining a diploma among students in deciles 2 through 5 is higher under the CAHSEE 1 regime. In contrast, the probability of graduation is lower under CAHSEE 2 in the bottom two deciles of the distribution.

FIGURE 1.8 HERE

The regression results of the difference-in-difference effects of the exit exams on overall graduation rates are presented in Table 1.8. The initial exam is associated with positive effects among students in deciles three, four and six, and further positive effects among the top achieving students. The motivational effects of CAHSEE 1 appear to be extending to the likelihood of graduation within five years. The findings suggest that the CAHSEE 1 not only can be associated with increases in student scores, but further increases in on-time and overall graduation rates. The positive effects among students in the 3\textsuperscript{rd} and 4\textsuperscript{th} decile of the achievement distribution vary

\textsuperscript{15}Graduation rate is the probability that a student obtains a high school diploma within 5 years of being enrolled in grade 9 for the first time.
between 1.7 and 1.8 percentage points. Given the observed positive effects in the top of the distribution in student test scores and grades in relation to CAHSEE 1, and the fact that the exam requirement was rescinded, it is possible that students learned more which in turn could have increased their graduation rates.

TABLE 1.8 HERE

The impacts of CAHSEE 2 differ from the effects of CAHSEE 1. CAHSEE 2 affects the probability of the overall graduation only for the bottom decile of the achievement distribution. For the lowest achievers CAHSEE 2 lowers the probability of graduation within 5 years by 25.1 percentage points, a 32% decrease from the pre-exam mean probability of graduation among the bottom decile. The findings are consistent with the theoretical predictions of the discouragement effect, concentrated at the bottom of the achievement distribution. For a significant fraction of low performers, CHASEE 2 proves to be a barrier to graduation. The analysis presented so far highlights the large negative effects of the exit exams among low achievers, investigating the effects of exit exams along the student achievement dimension. However, the analysis presented so far provides no insight into the potential effects associated with the overlap of material between the exam and courses, another dimension of heterogeneity which could potentially be affected by the introduction of high-stakes exams.

1.6.3 Exit Exam and Math Course Performance

An introduction of the exit exam should provide stronger incentive for students to learn the material most likely to be covered on the exam as well as in other graduation requirements. The math portion of the CAHSEE is designed to test ma-
terial covering grades 6 through 8, with topics in statistics, algebra, measurement and others. The most common courses taken in grade 10 are Algebra 1, High School Geometry and Algebra 2. The first course is the easiest, and the last, the most difficult one. The material content of the exit exam covers topics in Algebra 1, suggesting that the introduction of the exam should induce increased incentives to learn the Algebra 1 material. Given that Algebra 2 is a continuation of the first algebra course, the introduction of the exam might also induce increased incentives to learn the material covered in Algebra 2. In contrast, the exam does not extend to the topics covered in High School Geometry, but tests basic concepts in the subject, below the level of the high school course. Thus, one would expect that the alignment of incentives to study would be the lowest in Geometry.

The regression results of the triple difference model of the effects of exit exams on student achievement on mathematics test scores by math course taken in grade 10 are presented in Table 1.9. The first difference is the achievement before and after the introduction of the exit exam regime. The second difference is between the scores in grade 10 as compared to grades 6 through 8. The third difference is between the math courses taken in grade 10. Students in grade levels other than grade 10 are grouped according to what math course they take in grade 10. This way the comparison group encompasses a more appropriate group of students rather than the entire student population in middle school.

Regression results in Table 1.9 are suggestive of increased motivation to learn the algebra material translating into better performance. I do not model student selection into courses directly. I control for student position in the achievement distribution, which should be strongly correlated with course choices in grade 10. The investigation of the characteristics of the students enrolling in each course across the three graduation regimes suggests a stable selection process with the introduction of
the CAHSEE. The selection into courses requires further investigation to assure that the introduction of the exit exam does not affect the course enrollment process, which could bias the estimated results.

Table 1.9 suggests that the motivational effects among students enrolled in Algebra 1 might be mitigated by the discouragement effects found in the bottom of the achievement distribution, as the students in Algebra 1 are more likely to be low-achievers. The coefficients associated with the effects of both exit exams among Algebra 1 grade 10 students are all positive. However, the coefficients are statistically significant only with the inclusion of school specific time trends. CAHSEE 1 is associated with a 0.16 of a standard deviation increase in math scores among students in Algebra 1 in grade 10 when the regression accounts for differential school trends. CAHSEE 2 appears to be associated with increases in math scores by 0.258 of a standard deviation for students in the same course in grade 10.

In contrast, the most likely high achieving students, those enrolled in Algebra 2, experience increases in test scores associated with the introduction of the exit exams. CAHSEE 1 appears to increase math scores by 0.47 of a standard deviation among pupils in Algebra 2 in grade 10, controlling for school-specific time trends. CAHSEE 2 increases scores by 0.252 of a standard deviation for students in the same course in grade 10. The estimated total effects capture the potential overlap of material providing stronger incentives to learn, but also the motivational effects estimated in the upper part of the achievement distribution. Controlling for student position in the achievement distribution should in part mitigate potential issues of selection into courses.

For students enrolled in Geometry, the overlap of content with the exit exam should be the lowest. Furthermore the students in the middle of the achievement distribution would be expected to enroll in the course. The estimation results sug-
gest that CAHSEE 1 is associated with no net effect among the students enrolled in Geometry in grade 10. In contrast, CAHSEE 2 decreases math scores by 0.125 of a standard deviation for students enrolled in Geometry in grade 10, with the magnitude of the effect being significantly higher if the differential trends are excluded from the model. Finally, for students enrolled in other courses, potentially credit deficient and not on a regular academic schedule, exit exams are associated with decreases in student scores. The inclusion of school-specific time trends makes the estimated effects not statistically significant, which might be due to struggling students attending low-achieving schools. Thus, the controls for school-specific time trends might already be capturing the low levels of achievement. This group of students should experience the largest overlap of material between math courses and exit exam. Unfortunately, this is also the group which would be expected to be discouraged by the introduction of the exit exams. Therefore, the lack of statistical significant is not surprising, as the incentive effects would work in the opposite direction than the discouragement effects.

TABLE 1.9 HERE

The test score analysis seems to support the predictions of content overlap. The regression results investigating the effects of exit exams on math grades by math course taken in grade 10 are presented in Table 1.10. The introduction of the CAHSEE 1 regime does not seem to induce strong differential effects by course taken in grade 10. Only among the students enrolled in Algebra 2, the exam is associated with an up to a 0.18 increase in math grades, with most specifications significant only at 10%. The results suggest that the differential performance on standardized tests does not carry over to performance in the math courses themselves, under the CAHSEE 1 regime.
The results estimated in relation to the CAHSEE 2 exam indicate no significant differential effects associated with Algebra courses. However, the estimated coefficients are positive, suggesting that the increased incentives associated with the alignment of the exit exam with course content might be a factor increasing performance. In contrast, for students enrolled in Geometry and courses other than Algebra, the later exit exam regime is associated with negative effects on math grades. For students enrolled in Geometry in grade 10, the exam is associated with up to a 0.33 of a drop in math grades in grade 10. Among students enrolled in a math course other than Algebra or Geometry, the exam is associated with up to a 0.16 decrease in math grades when in grade 10, however accounting for school-specific time trends, the effect is not statistically significant. The analysis shows that the effects of exit exams on grades is not as pronounced as on test scores. Part of the explanation might be the cumulative grade-point average graduation requirements, which continue motivating students to do well in their courses.

The results presented in Table 1.9 and Table 1.10 are suggestive of increased incentives for learning among students for whom the exam overlaps with the course material. The regression results take into account differential achievement composition of the students in each course, by controlling for the student location in the grade 8 achievement distribution. Thus, the results presented should be viewed as a combination of the effects associated with the math course in grade 10 and student’s position in the initial achievement distribution.

The evidence presented thus far suggest that the introduction of the exit exam appears to lower achievement among low performing students, but the negative ef-
effects might be partially offset by the increase in the material overlap between the exam and student courses. The analysis suggest that the negative policy effects are concentrated among low achieving students.

1.7 Conclusions

Educational reform introducing high-stakes testing in California appears to have significant effects on student achievement. The analysis reveals that CAHSEE 1 induced strong incentive effects across the entire student distribution. In contrast, CAHSEE 2 is associated with negative effects towards the bottom of the achievement distribution. The analysis of student achievement, dropout and graduation rates align well with the theoretical predictions of potential effects associated with an introduction of test-based accountability. The analysis highlights the importance of the overlap of exam material with the course content and how the potential misalignment might negatively affect student performance on standardized tests and student coursework.

The discouragement effects detected in relation to CAHSEE 2 in most metrics of performance considered in the analysis suggest that the effects among students already struggling with educational requirements should be a serious concern for the policy makers. The exit exam introduced was designed to cover material well below the syllabus expectations of student performance in high school. In that context, the fact that the exam appears to have significant negative effects on graduation rates, graduation delays and lower performance on standardized tests suggests that educational deficiencies of students struggling in school might be a lot more severe than typically assumed, and the exam regime further exposed those problems. The remedial instruction associated with the exit exams administered in California typically targets students who already failed the exam once. The findings suggest that this
approach might be “too little, too late”, suggesting that pre-emptive remedial help might be better suited in this case.$^{16}$

The analysis presented in this paper suggests that the positive effects of test-based accountability were concentrated among the students exposed to the initial exit exam regime. The findings may be the result of relatively little knowledge about the exam difficulty, when it was first introduced. The subsequent decrease in motivational effects might be connected to students and teachers learning about the CAHSEE. The exam currently in place can be viewed as a minimum competency test and the motivational effects associated with it are very small. However, the exam still appears to induce discouragement effects among the low-achievers.

Given the lack of significant gains in achievement and statistically significant negative effects among the low performing students, the research suggests the need for debate on the exam and how the accountability system could be improved. The results indicate that with a careful consideration of material overlap, remedial instruction which addresses learning deficiencies early, exit exams can provide students with strong incentives to learn. The highlighted negative effects among low achieving students should be taken into account and be addressed in development of future educational accountability policies.

1.8 Tables

---

$^{16}$Betts, Zau, Zieleniak and Bachofer (2012) find that the remedial programs associated with CAHSEE had little effect on improving student outcomes and that the struggling students could be identified well in advance of the exam administration.
### Table 1.1: Timeline of the CAHSEE by Grade Level and School Year

<table>
<thead>
<tr>
<th>School Year</th>
<th>Grade Level</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997-1998</td>
<td>No Exam</td>
<td>No Exam</td>
<td>No Exam</td>
<td>No Exam</td>
<td></td>
</tr>
<tr>
<td>2001-2002</td>
<td>No Exam</td>
<td>CAHSEE 1</td>
<td>No Exam</td>
<td>No Exam</td>
<td></td>
</tr>
<tr>
<td>2002-2003</td>
<td>No Exam</td>
<td>CAHSEE 1</td>
<td>CAHSEE 1</td>
<td>No Exam</td>
<td></td>
</tr>
<tr>
<td>2003-2004</td>
<td>No Exam</td>
<td>CAHSEE 2</td>
<td>No Exam</td>
<td>No Exam</td>
<td></td>
</tr>
<tr>
<td>2004-2005</td>
<td>No Exam</td>
<td>CAHSEE 2</td>
<td>CAHSEE 2</td>
<td>No Exam</td>
<td></td>
</tr>
<tr>
<td>2005-2006</td>
<td>No Exam</td>
<td>CAHSEE 2</td>
<td>CAHSEE 2</td>
<td>CAHSEE2</td>
<td></td>
</tr>
</tbody>
</table>

### Table 1.2: Student Characteristics by Exam Type, Grade 6-12

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>No CAHSEE</th>
<th>CAHSEE 1</th>
<th>CAHSEE 2</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade-Point Average</td>
<td>2.704</td>
<td>2.763</td>
<td>2.755</td>
<td>2.741</td>
</tr>
<tr>
<td></td>
<td>(0.803)</td>
<td>(0.838)</td>
<td>(0.822)</td>
<td>(0.820)</td>
</tr>
<tr>
<td>Math GPA</td>
<td>2.225</td>
<td>2.218</td>
<td>2.202</td>
<td>2.211</td>
</tr>
<tr>
<td></td>
<td>(1.231)</td>
<td>(1.272)</td>
<td>(1.272)</td>
<td>(1.260)</td>
</tr>
<tr>
<td>ELA GPA</td>
<td>2.506</td>
<td>2.542</td>
<td>2.494</td>
<td>2.506</td>
</tr>
<tr>
<td></td>
<td>(1.142)</td>
<td>(1.173)</td>
<td>(1.183)</td>
<td>(1.170)</td>
</tr>
<tr>
<td>Math Score</td>
<td>-0.055</td>
<td>0.106</td>
<td>0.276</td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td>(0.982)</td>
<td>(1.005)</td>
<td>(1.155)</td>
<td>(1.086)</td>
</tr>
<tr>
<td>ELA Score</td>
<td>-0.015</td>
<td>0.176</td>
<td>0.392</td>
<td>0.225</td>
</tr>
<tr>
<td></td>
<td>(1.015)</td>
<td>(0.963)</td>
<td>(1.043)</td>
<td>(1.036)</td>
</tr>
<tr>
<td>English Learner</td>
<td>0.206</td>
<td>0.169</td>
<td>0.147</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>(0.405)</td>
<td>(0.375)</td>
<td>(0.354)</td>
<td>(0.374)</td>
</tr>
<tr>
<td>Absence Percentage</td>
<td>4.996</td>
<td>4.851</td>
<td>4.571</td>
<td>4.749</td>
</tr>
<tr>
<td></td>
<td>(6.082)</td>
<td>(5.684)</td>
<td>(6.055)</td>
<td>(6.001)</td>
</tr>
<tr>
<td>Observations</td>
<td>123,188</td>
<td>81,754</td>
<td>203,619</td>
<td>408,561</td>
</tr>
</tbody>
</table>

Table 1.3: Testing Parallel Trends, Using 1997-1998 to 2000-2001 Data

<table>
<thead>
<tr>
<th></th>
<th>Math Score</th>
<th>ELA Score</th>
<th>Math GPA</th>
<th>ELA GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time x Grade10</td>
<td>0.003</td>
<td>-0.001</td>
<td>0.005</td>
<td>-0.011</td>
</tr>
<tr>
<td>Time</td>
<td>0.034**</td>
<td>0.025**</td>
<td>-0.024**</td>
<td>0.010**</td>
</tr>
<tr>
<td>Grade10</td>
<td>0.061**</td>
<td>0.112**</td>
<td>-0.141**</td>
<td>-0.065**</td>
</tr>
<tr>
<td>Observations</td>
<td>194,056</td>
<td>192,606</td>
<td>210,358</td>
<td>217,032</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses.

** 1% p-level; * 5% p-level
Table 1.4: Triple Difference Estimate of the Impact of the CAHSEE on Low-Stakes Test Scores

<table>
<thead>
<tr>
<th></th>
<th>ELA Score</th>
<th>Math Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 1</td>
<td>0.095*</td>
<td>-0.204**</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 2</td>
<td>0.265**</td>
<td>-0.114**</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 3</td>
<td>0.245*</td>
<td>-0.130**</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 4</td>
<td>0.305**</td>
<td>-0.097**</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 5</td>
<td>0.394**</td>
<td>-0.040</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 6</td>
<td>0.372**</td>
<td>-0.050*</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 7</td>
<td>0.408**</td>
<td>-0.037</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 8</td>
<td>0.335**</td>
<td>-0.039</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 9</td>
<td>0.325**</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 10</td>
<td>0.272**</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.044)</td>
</tr>
</tbody>
</table>

Full Set of Controls ✓ ✓
Observations 410,473 408,259
R-squared 0.747 0.663

Standard errors in parentheses, clustered at the school level.
** 1% p-level; ** 5% p-level;
Table 1.5: Triple Difference Estimate of the Impact of the CAHSEE on Student Grades

<table>
<thead>
<tr>
<th></th>
<th>ELA GPA</th>
<th>Math GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 1</td>
<td>0.036</td>
<td>-0.136*</td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
<td>(0.053)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 2</td>
<td>0.144*</td>
<td>-0.053</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 3</td>
<td>0.182*</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 4</td>
<td>0.237**</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 5</td>
<td>0.173**</td>
<td>0.097*</td>
</tr>
<tr>
<td></td>
<td>(0.051)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 6</td>
<td>0.248**</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
<td>(0.063)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 7</td>
<td>0.318**</td>
<td>0.146**</td>
</tr>
<tr>
<td></td>
<td>(0.086)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 8</td>
<td>0.342**</td>
<td>0.186**</td>
</tr>
<tr>
<td></td>
<td>(0.063)</td>
<td>(0.048)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 9</td>
<td>0.349**</td>
<td>0.235**</td>
</tr>
<tr>
<td></td>
<td>(0.080)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>CAHSEE x Grd10 x Decile 10</td>
<td>0.280**</td>
<td>0.274**</td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>Full Set of Controls</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Observations</td>
<td>437,783</td>
<td>424,675</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.388</td>
<td>0.395</td>
</tr>
</tbody>
</table>

Standard errors in parentheses, clustered at the school level.

** 1% p-level; * 5% p-level;
Table 1.6: Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of Dropping Out by Decile of Initial Achievement (dp/dx)

<table>
<thead>
<tr>
<th>Decile:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1 x Decile</td>
<td>0.005</td>
<td>-0.005</td>
<td>0.001</td>
<td>-0.005</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.006)</td>
<td>(0.005)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>CAHSEE 2 x Decile</td>
<td>0.011*</td>
<td>0.004</td>
<td>0.010*</td>
<td>-0.005</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.003)</td>
<td>(0.006)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decile:</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1 x Decile</td>
<td>-0.011*</td>
<td>-0.003</td>
<td>0.000</td>
<td>-0.005</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.007)</td>
<td>(0.012)</td>
<td>(0.005)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>CAHSEE 2 x Decile</td>
<td>-0.008*</td>
<td>0.001</td>
<td>0.002</td>
<td>-0.011*</td>
<td>-0.012**</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.005)</td>
<td>(0.004)</td>
</tr>
</tbody>
</table>

Pseudo R-squared: 0.235  Obs: 48,000  Mean Dep. Var.: 0.042

Standard errors in parentheses, clustered at the school level.

* 1% p-level; 5% p-level;

Table 1.7: Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of On-Time Graduation by Decile of Initial Achievement (dp/dx)

<table>
<thead>
<tr>
<th>Decile:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1</td>
<td>-0.027</td>
<td>0.010</td>
<td>0.019**</td>
<td>0.019**</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.006)</td>
<td>(0.004)</td>
<td>(0.006)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>CAHSEE 2</td>
<td>-0.519**</td>
<td>-0.306*</td>
<td>-0.247*</td>
<td>-0.179</td>
<td>-0.182</td>
</tr>
<tr>
<td></td>
<td>(0.144)</td>
<td>(0.129)</td>
<td>(0.115)</td>
<td>(0.101)</td>
<td>(0.105)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decile:</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1</td>
<td>0.026**</td>
<td>0.015</td>
<td>0.023**</td>
<td>0.021**</td>
<td>0.029**</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.009)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>CAHSEE 2</td>
<td>-0.138</td>
<td>-0.210</td>
<td>-0.150</td>
<td>-0.153</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>(0.090)</td>
<td>(0.113)</td>
<td>(0.093)</td>
<td>(0.107)</td>
<td>(0.067)</td>
</tr>
</tbody>
</table>

Pseudo R-squared: 0.467  Obs.: 47,824  Mean Dep. Var.: 0.865

Standard errors in parentheses, clustered at the school level.

* 1% p-level; 5% p-level;
Table 1.8: Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of Graduation by Decile of Initial Achievement (dp/dx)

<table>
<thead>
<tr>
<th>Decile:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1</td>
<td>-0.030</td>
<td>0.007</td>
<td>0.018**</td>
<td>0.017**</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.006)</td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>CAHSEE 2</td>
<td>-0.251*</td>
<td>-0.135</td>
<td>-0.094</td>
<td>-0.055</td>
<td>-0.077</td>
</tr>
<tr>
<td></td>
<td>(0.125)</td>
<td>(0.096)</td>
<td>(0.078)</td>
<td>(0.058)</td>
<td>(0.066)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decile:</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1</td>
<td>0.020**</td>
<td>0.006</td>
<td>0.009</td>
<td>0.015</td>
<td>0.027**</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.012)</td>
<td>(0.009)</td>
<td>(0.008)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>CAHSEE 2</td>
<td>-0.053</td>
<td>-0.075</td>
<td>-0.077</td>
<td>-0.044</td>
<td>-0.024</td>
</tr>
<tr>
<td></td>
<td>(0.056)</td>
<td>(0.063)</td>
<td>(0.068)</td>
<td>(0.053)</td>
<td>(0.045)</td>
</tr>
</tbody>
</table>

Pseudo R-squared: 0.405  Obs.: 37,367  Mean Dep. Var.: 0.855

Standard errors in parentheses, clustered at the school level.
** 1% p-level; * 5% p-level;
Table 1.9: Triple Difference Estimate of the Impact of the CAHSEE on Math Scores by Math Course Taken

<table>
<thead>
<tr>
<th>Math Course Taken</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1 x Grd10 x Algebra 1</td>
<td>0.091</td>
<td>0.111</td>
<td>0.094</td>
<td>0.160*</td>
</tr>
<tr>
<td>CAHSEE 1 x Grd10 x Algebra 2</td>
<td>0.489**</td>
<td>0.481**</td>
<td>0.468**</td>
<td>0.470**</td>
</tr>
<tr>
<td>CAHSEE 1 x Grd10 x Geometry</td>
<td>-0.009</td>
<td>0.003</td>
<td>0.012</td>
<td>0.088</td>
</tr>
<tr>
<td>CAHSEE 1 x Grd10 x Other Courses</td>
<td>-0.134*</td>
<td>-0.127*</td>
<td>-0.133*</td>
<td>-0.047</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Algebra 1</td>
<td>0.122</td>
<td>0.081</td>
<td>0.088</td>
<td>0.258**</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Algebra 2</td>
<td>0.231**</td>
<td>0.169**</td>
<td>0.156**</td>
<td>0.252**</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Geometry</td>
<td>-0.302**</td>
<td>-0.334**</td>
<td>-0.302**</td>
<td>-0.125**</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Other Courses</td>
<td>-0.249**</td>
<td>-0.293**</td>
<td>-0.247**</td>
<td>-0.040</td>
</tr>
</tbody>
</table>

Student and School Characteristics | ✓      | ✓      | ✓      | ✓      |
Grade Level Fixed Effects           | ✓      | ✓      | ✓      | ✓      |
Year Fixed Effects                  | ✓      | ✓      | ✓      | ✓      |
School Fixed Effects                | ✓      | ✓      | ✓      | ✓      |
School-Specific Time Trends         | ✓      | ✓      | ✓      | ✓      |
Observations                         | 408,259| 408,259| 408,259| 408,259|
R-squared                            | 0.548  | 0.549  | 0.560  | 0.564  |

Standard errors in parentheses, clustered at the school level.
** 1% p-level; * 5% p-level;
Table 1.10: Triple Difference Estimate of the Impact of the CAHSEE on Math Grades by Math Course Taken

<table>
<thead>
<tr>
<th>Course Taken</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1 x Grd10 x Algebra 1</td>
<td>0.030</td>
<td>-0.051</td>
<td>-0.035</td>
<td>-0.023</td>
</tr>
<tr>
<td>CAHSEE 1 x Grd10 x Algebra 2</td>
<td>0.178**</td>
<td>0.129</td>
<td>0.120</td>
<td>0.125</td>
</tr>
<tr>
<td>CAHSEE 1 x Grd10 x Geometry</td>
<td>-0.078</td>
<td>-0.129</td>
<td>-0.123</td>
<td>-0.127</td>
</tr>
<tr>
<td>CAHSEE 1 x Grd10 x Other Courses</td>
<td>-0.009</td>
<td>-0.057</td>
<td>-0.063</td>
<td>-0.080</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Algebra 1</td>
<td>0.396*</td>
<td>0.292</td>
<td>0.282</td>
<td>0.322*</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Algebra 2</td>
<td>0.148*</td>
<td>0.076</td>
<td>0.074</td>
<td>0.119</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Geometry</td>
<td>-0.261**</td>
<td>-0.331**</td>
<td>-0.310**</td>
<td>-0.281**</td>
</tr>
<tr>
<td>CAHSEE 2 x Grd10 x Other Courses</td>
<td>-0.075</td>
<td>-0.158*</td>
<td>-0.157*</td>
<td>-0.117</td>
</tr>
</tbody>
</table>

Student and School Characteristics
Grade Level Fixed Effects ✓ ✓ ✓ ✓ ✓
Year Fixed Effects ✓ ✓ ✓ ✓
School Fixed Effects ✓ ✓
School-Specific Time Trends ✓
Observations 424,675 424,675 424,675 424,675
R-squared 0.341 0.342 0.355 0.358

Standard errors in parentheses, clustered at the school level.
** 1% p-level; * 5% p-level;
1.9 Figures

Figure 1.1: Correlations between SAT9 and CST Scores in Mathematics

Figure 1.2: Correlations between SAT9 and CST Scores in ELA
Figure 1.3: Student Scores and GPA in Grades 8 and 10 in Math and English by School Year

Figure 1.4: Exit Exam Effects on Test Scores by Decile of Achievement
Figure 1.5: Exit Exam Effects on GPA by Decile of Achievement

Figure 1.6: Dropout Rates by Decile and Exam Regime
Figure 1.7: On-Time Graduation Rates by Decile and Exam Regime

Figure 1.8: Graduation Rates by Decile and Exam Regime
1.10 Model of Exit Exams and Achievement

1.10.1 Student Achievement and Testing

I build on the theoretical models of academic achievement and standards in order to account for the possibility of multiple graduation requirements and characterization of the knowledge domain.\textsuperscript{17} Consider a hybrid human capital/signaling model in which school attendance leads to learning and adds to productivity. Each student has an initial level of ability \( a \), where \( a \) is a random variable on \([a, \bar{a}]\) with probability density and cumulative distribution functions \( f(a) \) and \( F(a) \), respectively. Students maximize utility with respect to leisure and wages \( U(L, w) \), where \( L \in [0, \bar{L}] \).

Following Betts (1998) academic achievement of student \( i \) \( (\pi_i) \) will be summarized through the education production function dependent on student’s ability and exerted effort, \( e \). The value of marginal product of worker \( i \) is directly proportional to worker’s academic achievement:

\[
VMP_i = \alpha \pi_i = \alpha g(\bar{L} - L_i, a_i) \quad \text{where} \quad g_1 > 0, \quad g_2 > 0, \quad g_{11} \leq 0, (1.4)
\]

\[ g_{22} \leq 0 \quad \text{and} \quad \alpha > 0, \quad \sum e_i = \bar{L} - L_i \]

Firms observe the academic standard each student meets and have the ability to distinguish between students who graduate from high school and those who do not. Firms use the graduation signal to set wages according to the expected productivity of each worker.\textsuperscript{18}

Student \( i \) is expected to meet an academic standard \( \pi_s \) and faces a set of graduation requirements which determine whether the student obtains a high school diploma or not. The student decides whether to exert effort in order to meet the grad-

\textsuperscript{17} I use models of academic standards introduced by Costrell(1994) and Betts (1998)

\textsuperscript{18} Following Betts(1998): \( w_1 = \frac{\int_{a}^{a^*} \alpha \pi f(a) da + \int_{0}^{a^*} g(0,a) f(a) da}{1 - F(a)} \) and \( w_2 = \frac{\int_{a}^{a^*} g(0,a) f(a) da}{F(a)} \)
uation requirements. Given the distribution of initial ability the student population can be divided into three main groups- students who do not meet the requirement, students who exert the effort in order to meet the requirement and students who meet the requirement without exerting any effort:

\[
\pi_i = \begin{cases} 
  g(0, a_i) < \pi_s & \text{if } a_i < a^* \\
  g(\bar{L} - L_i, a_i) = p_{i_s} & \text{if } a_i \in [a^*, a^{**}) \\
  g(0, a_i) > \pi_s & \text{if } a_i \geq a^{**}
\end{cases}
\]  

(1.5)

Consider the mechanism through which students decide what material to learn. Let N be the number of items in the knowledge domain covered by the school syllabus. The standard \(\pi_s\) will be defined as a subset \(N_Q\) of the knowledge domain expected to be mastered by a student in order to receive a diploma. The standard can be described as fraction \(Q\) of the items covered in school: \(\pi_s = QN\ s.t.\ Q \in (0, 1]\).

At grade level \(p\), a student is expected to meet a grade specific requirement which encompasses a fraction of \(Q_pN\), the material covered at the grade \(p\) level. At each grade level only \(M\) items can be monitored reliably and the state can announce the parts of the syllabus to be fair game for requirement testing as \(q_pN\), where \(q_p \in [0, Q_p]\). Each item covered in grade \(p\) has a probability of \(\frac{M}{q_pN}\) of being tested\(^{19}\). The school requires the student to be able to answer correctly fraction \(s\) of the items on the test. Define the loss of utility from exerting effort to learn item \(j\) as:

\[
d(e_{ij}, a_i) = e_{ij}^*(a_i) \frac{\partial U(L_i, w_1)}{\partial L_i} \bigg|_{L=L_i}
\]

and the punishment of not learning the item as the utility difference between meeting the standards and exerting effort, and not meeting the standards, but exerting no

\(^{19}\)The formulation follows Lazear(2006) decision framework for when a student might learn an item
effort:

\[ K_{\pi_s} = U(L_i, w_1)|_{L_i} - U(L, w_2) \]  

(1.7)

Thus, student i learns item j if the loss of leisure resulting from learning the item, at student i’s ability level, is smaller than the punishment K for not learning the item:

\[ d(e_{ij}, a_i)|_{L_i} \leq K_{\pi_s} \frac{sM}{q_p N} \]  

(1.8)

Student i meets the grade specific requirements if:

\[ R^p_i = 1 \left[ \sum_{j=1}^{M} \left[ d(e_{ij}, a_i) \leq K \frac{sM}{q_p N} \right] \geq sM \right] \]  

(1.9)

I denote \( N_{q_p} \) as the subset of items fair game to be monitored from the grade p syllabus (\( N_{q_p} = q_p N \)). The characterization of the academic standard and of the grade specific requirements further assumes that \( N_Q \subset \bigcup_p N_{q_p} \). The positive probability of being monitored for all items in \( N_{q_p} \) implies that students might learn more than just the material which ends up being monitored if the student has sufficiently low cost of learning.

1.10.2 Exit Exams and Distribution of Achievement

Assume that an exit exam is introduced in grade p. The exam consists of C questions and covers a subset of the material required for graduation. Students are expected to know fraction s of the exam material, therefore the probability of an item appearing on the exit exam is \( \frac{sC}{q_c N} \). The student in grade p is required to satisfy the grade specific requirements and to pass the exit exam in order to graduate.

First, assume that the exit exam does not change the academic standard \( \pi_s \), but it requires additional effort on the part of the student, \( c^e_i \). Thus, the distribution
of student education achievement can be characterized as:

$$
\pi'_i = \begin{cases} 
  g(0, a_i) < \pi_s & \text{if } a_i < a^* \\
  g(0, a_i) < \pi_s & \text{if } a_i \in [a^*, a'] \\
  g(\bar{L} - L_i - e_i^{ex}, a_i) = \pi_s & \text{if } a_i \in (a', a'') \\
  g(0, a_i) > \pi_s & \text{if } a_i > a''
\end{cases}
$$  \hspace{1cm} (1.10)

Students at the bottom of the distribution do not change their behavior, as they fail to meet the standard and exert no effort. Students in $[a^*, a']$ interval would have met the academic standard in absence of the exit exam, but with the introduction of the exam they choose to exert no effort. Performance of those student should be lower than comparable students who do not have to pass an exit exam. Students in $(a', \bar{a}]$ continue meeting the standard and for them their achievement should be comparable to the students who were never required to pass an exit exam.

Second, consider a case where the introduction of an exit exam coincides with an increase in the academic standard, $\pi_s \rightarrow \pi'_s$. The average level of effort exerted by students who continue meeting the standard should increase. Student achievement across the ability distribution can be described as:

$$
\pi'_i = \begin{cases} 
  g(0, a_i) < \pi'_s & \text{if } a_i < a^* \\
  g(0, a_i) < \pi'_s & \text{if } a_i \in [a^*, a\dagger] \\
  g(\bar{L} - L_i - e_i^{ex}, a_i) = \pi'_s & \text{if } a_i \in (a\dagger, a^{\dagger\dagger}] \\
  g(0, a_i) > \pi'_s & \text{if } a_i > a^{\dagger\dagger}
\end{cases}
$$  \hspace{1cm} (1.11)

Similarly to the previous case, there should be no change in achievement for students in $a_i < a^*$, and the achievement of students in $[a^*, a\dagger]$ should decrease. In contrast, for students who meet the new academic standard and exert effort $(a_i \in (a\dagger, a^{\dagger\dagger})$, achievement should increase as compared to similar student facing an old academic
standard and not required to pass an exit exam. Students at the top of the achievement distribution should remain unaffected as the new standard continues to be below their academic level. This assertion is based on the idea of no gain from increasing achievement if the student meets the graduation requirements. In the real world we would expect that top performers might be induced to increase achievement in order to further distinguish themselves and signal their high quality to prospective colleges.

1.10.3 Exit Exams and Incentives to Learn

After characterizing the distributional effect of exit exams, consider the individual incentives associated with an introduction of an exit exam on grade specific performance and testing. Conditional on meeting the graduation requirements, students will make decisions regarding learning depending on the severity of the repercussions of not learning a particular item:

\[
d(e_{ij}, a_i) \leq \begin{cases} 
K_{\pi s} \frac{sM}{q_pN} & \text{if } j \in \{N_Q \setminus N_{Qc}\} \\
K_{\pi s} \frac{sC}{q_cN} & \text{if } j \in \{N_{Qc} \setminus N_Q\} \\
K_{\pi s} \left[\frac{sM}{q_pN} + \frac{sC}{q_cN} - \frac{s^2CM}{q_cq_pN}\right] & \text{if } j \in N_Q \cap N_{Qc} 
\end{cases}
\] (1.12)

Equation 1.12 suggest that students have the highest incentives to learn items with a positive likelihood to be asked in both grade specific requirements and the exit exam. If the student has to multi task, he or she would be expected to focus on tasks with the highest incentives first. The overlap of course content and the material covered on the exam should have a positive effect on the learning incentives. Students enrolled in classes more aligned with the CAHSEE should exhibit a differentially higher achievement on standardized tests that students for whom the course content does not have a large overlap with the exam.
1.11 Acknowledgments

I thank my thesis committee: Prof. Julian R. Betts, Prof. Craig McIntosh, Prof. Karthik Muralidharan, Prof. Prashant Bharadwaj, and Prof. Amanda Datnow. I would also like to thank Prof. Julie Cullen, Prof. Gordon Dahl, Prof. Roger Gordon, Michael Kuhn, Dallas Dotter and seminar participants at UC San Diego for all the valuable comments and suggestions. I would like to extend further thanks to Karen Volz Bachofer and Andrew Zau for the technical and institutional support and to the San Diego Unified School District for the data used in the analysis.
Chapter 2

Effects of Test Failure on Academic Achievement in the Context of Student Accountability: A Regression Discontinuity Analysis.

Jedrzej [Yendrick] Zieleniak\textsuperscript{1} and Julian R. Betts\textsuperscript{2}

\textsuperscript{1}University of California, San Diego, Department of Economics, 9500 Gilman Drive, La Jolla, CA 92093, USA. E-mail: jzieleni@ucsd.edu. Web page: http://econweb.ucsd.edu/~jzieleni.

\textsuperscript{2}University of California, San Diego, Department of Economics, 9500 Gilman Drive, La Jolla, CA 92093, USA. E-mail: jbetts@ucsd.edu. Web page: http://econweb.ucsd.edu/~jbetts/.
2.1 Abstract

We investigate the effects of initial failure on the California High School Exit Exam on student outcomes. We find that for students close to the passing cutoff, the initial failure does not affect student achievement while in high school. However, failing either component of the test increases the likelihood of enrollment in remedial classes the following year. Furthermore, we find that scoring just below the passing cutoff on the exit exam in grade 10 does not affect students’ likelihood of dropping out, or graduating with a diploma. An analysis of postsecondary outcomes shows that conditional on graduating from high school, initial CAHSEE failure has no effect on 2-year college enrollment or the likelihood of transfer to a 4-year university. We find evidence of increased probability of enrollment in a 4-year college, conditional on graduating, associated with the exam failure. Our analysis shows that the California High School Exit Exam does not pose a significant hindrance to student achievement for students close to the passing cutoff.

*JEL Classifications:* I210; I240; J240

*Keywords:* Education Policy, Exit Exams, Accountability, Learning, Human Capital, Regression Discontinuity
2.2 Introduction

Exit exams are a common student accountability tool used in many educational systems. The proliferation of this form of high-stakes testing in the U.S. is driven by the passing of the No Child Left Behind Act of 2001. As of 2012 more than half of the US states require high schools students to pass exit exams (Reardon, Arshan, Atteberry, and Kurlaender (2010)). Tests measuring student achievement are designed to hold students accountable to a specific academic standard or to benchmark the minimum proficiency at a given education level. In theory tests provide a uniform tool for higher education institutions to distinguish between students who meet the standard and those who do not (Becker (1982)). They further prove to be useful when individual ability is difficult to observe and employers have to rely on proxies to determine worker’s productivity in the initial employment period. The signal of passing a high school exit exam and graduating might provide the firm with information about how valuable the new hires are and how to set wages (Costrell (1994), Betts (1998)). This is only possible if uniform and comparable tests are administered to the entire student population. The market related motivation for uniform standards provides strong incentives for students, which should lead to increased achievement, at least among high school graduates.

The impacts of an exit exam depend on whether the exam is pitched at a low level (“minimum competency”) or at a high level. Theory suggests that introduction of a minimum-proficiency standard does not increase the threshold of material required to graduate. A minimum competency exam might be the first step in improvements of the educational system. The exam might force educators to standardize the level of instruction across schools. Under this scenario, the content of the exit exam to be introduced in the school system should be pitched at a level testing the minimum skills expected of all the students at a given education level. Alternatively,
introduction of an exit exam designed to test grade appropriate material would be expected to produce significantly different results\(^3\).

The proliferation of exit exams as accountability tools motivates a research agenda on the effects of exit exams on student outcomes. The main research focus has been on dropouts and graduation rates due to concerns that exit exams will encourage some students to leave school early\(^4\).

But the pass/fail character of exit exams also makes them suitable for a regression discontinuity analysis, allowing for the identification of the local treatment effects around the passing cutoff. Despite the proliferation of exit exams and other forms of high-stakes testing, the number of empirical studies investigating the effects of failure using the RD framework is quite limited.

The empirical RD studies find both positive and negative effects on student outcomes, depending on the context. Jacob and Lefgren (2004) find that failure on standardized tests and the remedial summer instruction increases student achievement among third graders but has no effect among sixth graders in the Chicago Public Schools in the period of 1993-1994 to 1998-1999. The findings can be interpreted as indicating the achievement gains possible with a well designed remedial education program.

In the context of high school exams Ou (2010) finds that failing an exit exam in the state of New Jersey increases the probability of dropping out among racial

\(^3\)The models of Costrell (1994) and Betts (1998) predict that students with achievement levels far below the standard would reduce their effort to zero, while those who were able to meet the standard would increase their effort. Introduction of a difficult exit exam would lower graduation rates due to some students reducing their effort and failing to meet the academic requirements. Their models assume that the exit exams are testing student knowledge at a reasonably high level.

\(^4\)Several studies have attempted to estimate the overall effects of exit exams on the student population. Some studies have focused on comparison of graduation patterns across states (Greene and Winters (2004), Warren and Edwards (2005), Jacob (2001) and Warren et al. (2006) to name a few) or school districts (Griffin and Heidorn (1996)). A small number of studies have used a within district difference in difference framework to estimate total effects of exit exams, such as Jacob (2005) or ?
minorities and economically disadvantaged students. Similarly, Papay, Murnane, and Willet (2010a) find negative effects associated with initial failure on the exit exam in the population of high school students in the state of Massachusetts who were in grade 10 in 2004. The authors find that the low-income urban students who just failed the mathematics examination had an 8 percentage point lower probability of graduating from high school than comparable students who barely passed. Negative effects of failure were also detected in student academic trajectories while in high school. Ahn (2014) finds that failure on high-stakes exams in North Carolina reduced the likelihood of enrollment in college track curriculum among 9th graders. Finally, Reardon, Arshan, Atteberry, and Kurlaender (2010) find no effects on student achievement, course-taking, persistence, or probability of graduation in California.

In this paper we employ a regression discontinuity design to estimate the causal effects of the failure on the California High School Exit Exam (hereafter referred to as the CAHSEE) in grade ten on student achievement in grade eleven, on the probability of dropping out, and on the probability of graduating from high school. Furthermore, we investigate the effects of the initial failure on the CAHSEE on the probability of postsecondary enrollments. We find no effect of failing either of the components of the CAHSEE in grade ten on most of these academic outcomes.

Students are permitted up to six attempts at taking the exam, over the course of grades ten through twelve. We focus our attention on the CAHSEE results from grade 10. By state law all high school students are mandated to take the exit exam at that grade level for the first time. Failure rates on the CAHSEE in grade 10 are relatively high with roughly 30% of students in our data failing the exam. Any attempts to model CAHSEE results in later grades potentially suffer from selection bias, as only the students who failed the test in the past are required to retake it.

Our study adds to the investigation of the effects of failure on the CAHSEE
conducted by Reardon, Arshan, Atteberry, and Kurlaender (2010). First, the authors had access to the four largest school districts in California, with data spanning five cohorts. In our case, we carry out the analysis in the second largest school district in California, but with seven cohorts in our panel. Second, their analysis focuses on student outcomes at high school level and omits measures such as math test scores and GPA, due to endogenous course selection by students. We analyze math low-stakes test scores and GPA and control for the courses taken to mitigate the issues of course selection. Third, our paper is the first one to analyze the effects of initial failure on postsecondary outcomes. We further add to the previous studies by considering the effects of exam failure on enrollment in remedial courses.

Ours is the first paper to investigate the effects of failure on likelihood of enrollment in remedial courses in high school. As such the results we report should be viewed as the net effect of failure, including the remedial resources directed towards the students after the first attempt at the exit exam. We find that barely failing either component of the CAHSEE increases the likelihood of enrolling in preparatory classes in the failed subject by 2-5.5% in Math and by 1.6-2.9% in English Language Arts (ELA), as compared to the students who barely passed the given component of the exam. For the remainder of the measures of student achievement while in high school, initial failure on the CAHSEE does not induce any effects on performance. The empirical analysis is consistent with theory, in the case where the exam serves the purpose of the introduction of a minimum competency standard.

It is possible that initial failure on the exam might change the academic trajectory of a student by shifting effort away from typical subjects and towards remedial courses, affecting postsecondary education decisions. We analyze probabilities of enrollment at a two-year college, and at a four-year university, conditional on graduating from high school, and the likelihood of transfer from a 2-year to a 4-year college, con-
ditional on 2-year college enrollment. We find that failing one of the components of the CAHSEE in grade 10 in general has no effect on the likelihood of enrollment in a 2-year college, or the likelihood of transfer. However, among students who failed CAHSEE Math in grade 10, those who also failed CAHSEE ELA have a 4.4% higher likelihood of enrolling in a 4-year college than those who initially passed ELA but failed Math. The results are stable across different bandwidths and suggest that the additional attention towards students at risk might be boosting student academic trajectories.

The remainder of the paper is organized as follows. Section 2 discusses the institutional context of the California High School Exit Exam. Section 3 provides the conceptual framework guiding the analysis, presents the data, and provides descriptive statistics. Section 4 presents the results. Section 5 concludes.

2.3 The California High School Exit Exam

The California High School Exit Exam is the only component of the education accountability system in the state with direct consequences for students. Each student is required to pass both portions of the exam in order to graduate from high school with a diploma. The development of the test was authorized by the California Education Code Section 60850(a), enacted in 1999. The test is intended to improve student achievement and to guarantee grade level competency in reading, writing and mathematics among high school graduates (California Department of Education 2011). The test was planned to be a binding requirement starting with the graduating class of 2004. However, low passage rates by the summer of 2003 forced the State Board of Education to suspend its implementation. Subsequently, the test was altered, with the reduction of the length of the ELA component and the simplification of the Math portion of the exam. The new version of the test was announced as a
requirement starting with the class of 2006 (Zau and Betts (2008)).

The initial introduction of the CAHSEE into the educational system resulted in the graduating classes of 2004 and 2005 taking the test in grade 10 under the conviction that passing the exam would be a graduation requirement for them. The summer 2003 decision to suspend the requirement implies that for those two cohorts initial failure on the exam should not carry any consequences. In contrast, starting with the graduating class of 2006, successfully passing the new version of the CAHSEE remained a requirement and initial failure on the exam might have an effect on student performance both in high school as well as on their future academic choices. We control for the initial version of the exam in the analysis to make sure that potentially different effect of failure on the early version of the exam does not skew our main estimates. In the regression tables we present the average effects of initial failure, as the differential effects associated with the first version of the exit exam are not statistically significant and do not affect the main analysis.

The CAHSEE consists of two portions, English Language Arts (ELA) and mathematics. The ELA test is aligned with the California academic content standards through grade ten, while the math exam is aligned with the California mathematics academic contents standard through the first part of Algebra I, which is typically taught in grade eight or nine. Overall, the math portion of the exam is designed to test skills covered in grades six to eight. Thus, in the context of being a high school exit exam the CAHSEE can be viewed as testing minimum competencies expected of graduates, and not an exam pitched at a very high level.

The ELA test consists of multiple choice questions and a written response. The test covers reading comprehension, analysis of literary texts and information as well as a written response to a writing prompt, literature, or an informational passage. The math component of the exam consists entirely of multiple choice questions
which cover statistics, data analysis and probability, number sense, measurement and geometry, mathematical reasoning, and algebra (California Department of Education 2011).

The passing score on each test is 350 points, or approximately 55 percent of items correct in ELA and 60 percent of items correct in math. If a student passes one of the two subject tests, but fails the other, he/she is only required to retake the failed test. The students are mandated to take the CAHSEE for the first time in grade ten. Should they be required to retake it, they have two opportunities in grade eleven and in the case of San Diego Unified School District up to three attempts in grade twelve. The fact that only students who fail a component of the exam have to retake it implies that there is a strong selection among those who take the exam after grade ten. The inclusion of results from the subsequent waves of the test would contaminate the analysis by introducing a strong selection bias. The initial attempt at taking the CAHSEE should provide the students with the most information and the strongest signal regarding their position in relation to the standard and therefore should have the strongest effect on student performance.

2.4 Conceptual Framework and Data

It is difficult to obtain an estimate of the causal effect of the impact of CAHSEE on all students due to unobserved student heterogeneity. Rather our goal is to estimate the causal impact of failing the CAHSEE for a more narrow group of students, those near the passing score cutoff. The students within a narrow margin of the standard can be considered as similar to one another. For those students the standard, at the time of taking the test, is a constraint which can provide them with valuable information regarding their effort and achievement. For students just above the passing cutoff, the exam is a signal that they passed by a narrow margin. For
those just below the passing score, the results are an indication that they are close to meeting the standard, but they need to increase their effort in order to meet the requirements set out by the school system.

Whether failure induces greater effort, no change in effort, or lower effort due to a discouragement effect, is an empirical question. In all cases, the exam is a valuable source of information which might induce a significant change in academic achievement. Furthermore students just below the passing cutoff may or may not receive remedial support. They might have the highest marginal returns to increased academic inputs and therefore they could receive remedial help. However, they might not receive additional attention if they are viewed as perfectly capable of reaching the standard without additional help.

2.4.1 RD Design for Student Achievement

We model student academic achievement after the initial CAHSEE attempt as a function of student’s past achievement ($A_{i0}$), a vector of school inputs ($Sch_{it}$), student’s own effort ($Ef_{it}$) and remedial help a student might receive as a result of initially failing the exit exam ($R_{it}$). We assume that achievement at time $t+1$ depends on the effort, school resources and remedial help in the previous period (the period leading to the current round of assessments). Let the academic achievement of student $i$ in school $s$ the year after the student takes CAHSEE for the first time ($t+1$) be described by the following model:

$$Y_{is(t+1)} = \alpha + \beta A_{i0} + \Gamma Sch_{it} + \delta Ef_{it} + \lambda R_{it} + \varepsilon_{is(t+1)} \quad (2.1)$$

School resources, student effort and remedial help after the first attempt at the exit exam might all depend on student and school’s beliefs about student ability level, and
on student’s initial CAHSEE score. The perceptions about student ability level should be strongly correlated with student past performance captured by the grade-point average (GPA\textsubscript{i0}) and low-stakes test scores (test\textsubscript{i0}), and should further be affected by the results of the first attempt at the CAHSEE (Ex\textsubscript{it}). We can re-write Equation 2.1 as:

\[ Y_{is(t+1)} = \alpha + \beta A_{i0} + \Gamma Schl_{it} (GPA_{i0}, test_{i0}, E_it) + \delta Ef_{it} (GPA_{i0}, test_{i0}, E_it) + \lambda R_{it} (GPA_{i0}, test_{i0}, E_it) + \varepsilon_{is(t+1)} \] (2.2)

While in general investments in student education are determined in part by underlying student ability and school resources, the remedial instruction and resources aimed at bringing the student on track towards completing graduation requirements should have a probability distribution with a discrete jump for the CAHSEE scores just below c=350. The modeling of the remedial help with the recognition that high ability students might test poorly on their first attempt, or that schools might have limited remedial resources, implies that the discrete jump in probability associated with receiving remedial help below the passing cutoff does not equal to one across all students just below the cutoff.

The proposed framework allows us to compare the right and left limits of the expected academic achievement after the first attempt at CAHSEE. Using (2.2) the difference in limits from below versus above can be written as:

\[
\lim_{Ex_{it} \uparrow c} \mathbb{E} [Y_{is(t+1)}|Ex_{it}] - \lim_{Ex_{it} \downarrow c} \mathbb{E} [Y_{is(t+1)}|Ex_{it}] \\
= \Gamma \left[ \lim_{Ex_{it} \uparrow c} Schl_{it} (GPA_{i0}, test_{i0}, E_it) - \lim_{Ex_{it} \downarrow c} Schl_{it} (GPA_{i0}, test_{i0}, E_it) \right] \\
+ \delta \left[ \lim_{Ex_{it} \uparrow c} Ef_{it} (GPA_{i0}, test_{i0}, E_it) - \lim_{Ex_{it} \downarrow c} Ef_{it} (GPA_{i0}, test_{i0}, E_it) \right] \\
+ \lambda \left[ \lim_{Ex_{it} \uparrow c} R_{it} (GPA_{i0}, test_{i0}, E_it) - \lim_{Ex_{it} \downarrow c} R_{it} (GPA_{i0}, test_{i0}, E_it) \right] \] (2.3)
At the limit, the difference in outcomes between the students just below and those just above the passing cutoff can be associated with the discrete jump in student effort, school resources, and remedial help devoted towards the student resulting from initial failure. Given that we don’t observe student effort, data on school resources is not available, and information regarding remedial help is limited, the regression discontinuity analysis can be viewed as the estimation of the total composite effect of initial failure on the CAHSEE.

We attempt to unpack the black box of the total effects associated with initial exam failure using data on exam preparatory courses. The San Diego Unified School District offers an array of CAHSEE preparatory classes. If the post-exam investments among those who barely failed on their first attempt include preparatory courses, then we would expect that the empirical analysis should show a higher probability of prep course enrollment just below the cutoff. In the case of all other outcomes the coefficients on the exam failure should be viewed as the overall effect of exam failure, including the remedial help as well as the psychological effects associated with motivation or discouragement.

2.4.2 Empirical Design

We use the regression discontinuity design in order to identify the role of academic inputs and changes in the levels of effort, following the framework of Imbens and Lemieux (2008). We formalize the empirical model using the following general specification, used for different outcome variables $y_{ijt}$ for student $i$ in school $j$ at time $t$, and different subsamples of the student population:

$$
y_{ijt} = \beta_0 + \beta_1 \mathbb{1}[E_{x_{ijtz}} < 0] + \beta_2 \mathbb{1}[E_{x_{ijtz}} < 0] \times E_{x_{ijtz}} + \beta_3 \mathbb{1}[E_{x_{ijtz}} \geq 0] \times E_{x_{ijtz}} + \delta_j + \delta_t + \gamma X_{ijt} + \sigma S_{jt} + \varepsilon_{ijt}
$$

(2.4)
where $Ex_{ijtz}$ is a standardized CAHSEE score of student i in grade ten on component $z$ of the exam, $z \in \{\text{ELA, Math}\}$. The CAHSEE scores used in the analysis are standardized by subtracting the passing minimum score of 350 from the student’s score. $1[Ex_{ijtz} < 0]$ is an indicator function equal to one if student i obtained a standardized score on component $z$ less than the cutoff of zero, $\delta_j$ is a school fixed effect, $\delta_t$ is a year fixed effect, $X_{ijt}$ is a vector of student characteristics, $S_{jt}$ is a vector of the school time-varying characteristics and $\varepsilon_{ijt}$ is the error term. We introduce the student and school characteristics, as well as differential slopes on both sides of the passing cutoff, in order to control for student heterogeneity as we move away from the cutoff score and the limits described in Equation (2.3).

The coefficient of interest is $\beta_1$, which captures the local average treatment effect of failing the CAHSEE at the point in the test score distribution where the passing score cutoff falls. Our results do not generalize to the entire student population, but rather focus on the subset of students in the neighborhood of the minimum passing score. The coefficients $\beta_2$ and $\beta_3$ allow for differential slope with respect to the CAHSEE score on the different sides of the passing cutoff. This approach to the treatment of the test score allows for independent estimation of the effect of the distance from the cutoff on student outcomes. $\beta_2$ and $\beta_3$ provide information on how student outcomes change, as we move farther away from the cutoff to the left and to the right respectively.

### 2.4.3 Bandwidth Selection

We use a range of four bandwidths around the passing cutoff in order to investigate whether the results obtained close to the cutoff can be extended to a wider group of students. The smallest bandwidth used in the analysis includes students within 5 points of the passing cutoff, a bandwidth of 10 points. Small bandwidth
should guarantee that the students included in the regression analysis should be ob-
ervationally similar to each other, regardless on which side of the cutoff they happen
to be. A ten point window, however, implies that only a small number of observa-
tions would be used in the analysis. Thus, we use wider bandwidths, each increasing
outwards in size by 10 points and ranging between 10 and 40 points, or 5 to 20 points
from the passing cutoff respectively.

An alternative approach would be to use optimal bandwidth selection meth-
ods such as the cross validation method, the second-generation plug-in bandwidth
selection approach proposed by Calonico et al. (2012) (CCT), or the Imbens and
Kalyanaraman (2012) (IK) consistent, robust bandwidth estimator. All three meth-
ods are proposed to provide an optimal bandwidth for the regression discontinuity
analysis. However, these methods provide a different set of bandwidths for each out-
come and each subsample in the analysis, making keeping track of the bandwidth
and the results more difficult. In general, the optimal bandwidth selection methods
suggested bandwidths between 10 and 74 points. The coefficients estimated using op-
timal bandwidth selection methods are similar to those obtained using the simplified
bandwidth selection procedure described above. We report the coefficients from the
simplified procedure as they provide the same set of bandwidth for each outcome.

2.4.4 Data

We use administrative data from San Diego Unified School District for the
first seven graduating classes which were required to take the CAHSEE. For the first
two classes the exit exam requirement was suspended when the students from the
first two cohorts finished grades 10 and 11 respectively. In the regression framework
we control for the fact that the first two cohorts might be affected differently by the
initial CAHSEE. In contrast, the later cohorts were required to pass the exam in order
to obtain a high school diploma. The data includes the graduating classes of 2004 through 2010. The data are distributed relatively evenly across the cohorts with a total of 66,609 observations. The subsamples used in the analysis, around the passing cutoffs, are significantly smaller.

Our data set contains records of student characteristics such as gender, ethnicity, age as well as English learner status and special education status. Furthermore, the data set contains information pertaining to student academic performance: California Standards Test (CST) scores in Math and English Language Arts from grade 6 to grade 11, Grade Point Average, CAHSEE Scores in grades 10, 11 and 12 as well as courses taken and grades obtained by the students in Math and English Language Arts during their high school career. We recognize that the CST is not a vertically scaled test. Thus, a direct comparison of test performance across grades using the test scores is not an appropriate approach.

In order to make test performance comparable we standardize scores within grade, within test type, across years, with a zero mean and variance equal to one in order to place student performance on a common scale by grade-level. In the case of ELA all students at any grade level in a given year take the same CST test, however, in math, students have a choice of which math course to take in grades 8 and above, each linked to a specific CST test. Thus the standardization of the scores within a test type, with further controls for the type of test taken addresses the issue of the tests being of varying difficulty in mathematics. The additional controls provide a more precise comparison group for each student, as they effectively group students according to the courses taken and compare them to the most relevant group of peers.

Failure on the CAHSEE on the first attempt in grade 10 can be viewed as a form of treatment. The regression discontinuity framework assumes that students within a narrow window of the passing cutoff score should be identical on observable
and unobservable characteristics. Naturally, when we extend that window to include a larger number of observations and we move further away from the cutoff, we expect larger differences in the observable student characteristics. We pick a window of 10x10 points around the passing cutoffs in Math and ELA in order to capture a significant number of students to be compared by whether they passed or failed either component of the exam. The window chosen to present baseline student characteristics captures approximately 4 percent of the student population.

TABLE 2.1 HERE
TABLE 2.2 HERE

Table 2.1 summarizes the means of student characteristics by subgroup, while Table 2.2 presents the t-tests of equality of means between groups for continuous variables and the Pearson Chi-Squared tests for categorical variables. Among the students who failed (passed) CAHSEE Math in grade 10, the difference in the Grade 9 CST ELA scores between those who passed and those who failed CAHSEE ELA in grade 10 is statistically significant at the 1% level, with the t-statistic equal to 5.519 (5.126). Similarly, among the students who failed (passed) CAHSEE ELA in grade 10, there is a difference in CST Math scores in grade 9 between those who passed and those who failed CAHSEE Math in grade 10, with a t-statistic of 2.354 (4.364), significant at the 1% level.

The differences in GPA are also significant for two comparison groups. Among students who failed CAHSEE Math in grade 10, the difference in GPA between students who failed and those who passed CAHSEE ELA is significant at the 5% level, with a t-statistic of 1.658. Similarly, among students who passed CAHSEE ELA in grade 10, the difference in GPA in grade 9 between students who passed and those
who failed CAHSEE Math in grade 10 is also significant at the 5% level. The percentage of days absent among students who passed CAHSEE ELA in grade 10 is statistically higher among the students who failed math as compared to those who passed it on their first attempt. Additionally, for students who passed CAHSEE Math in grade 10, there is a larger fraction of English learners who failed ELA in grade 10, as compared to those who passed ELA (Table 2.2 column 2, row 5). The difference is significant at the 1% level. Finally, for students who passed CAHSEE Math in grade 10, the fraction of Hispanic students is higher among those who failed CAHSEE ELA in grade 10 as compared to those who passed the ELA exam.

The baseline characteristics of the students in the ten point neighborhood of the passing cutoff reveal that the students in the four subgroups in the analysis differ on baseline characteristics. We realize that the students within 10 points from the passing cutoff are expected to be different from each other by subgroup. Thus, we include student characteristics in all of our regressions in order to control for the potential heterogeneity across students.

2.5 Empirical Results

We divide the analysis into three distinct sections. First, we investigate the effects of failing either component of the CAHSEE on the probabilities of enrolling in remedial classes in the subsequent years. Our analysis is the first one to investigate the effects of failure on remedial resources. The analysis provides insight into the remedial help given to students after their initial CAHSEE attempt. While the approach is not exhaustive in terms of the analysis of investments made after the initial CAHSEE attempt, the pattern of course taking should shed light on whether the initial failure on the exit exam changes academic trajectories for students who scored just below the cutoff.
Second, we carry out the empirical analysis focusing on six high school outcomes: the Grade-Point Average in grade 11, the CST English Language Arts score in grade 11, the CST Math score in grade 11, the probability of dropping out, the probability of graduating with a diploma two years after taking the CAHSEE for the first time, and the probability of graduating from high school within six years of enrolling in grade 9 for the first time. The analysis of dropouts and graduation rates is typically the main focus of analysis of initial failure on exit exams. However, we enrich the typical approach by analyzing more immediate outcomes like the GPA and test performance in grade 11. The rationale for the investigation of these medium-term outcomes stems from the theoretical framework presented in section 3. The remedial help potentially provided to students immediately after their initial CAHSEE failure could translate into academic gains associated with the material covered in school. Thus, the potential cognitive gains associated with remedial help could go undetected in long term outcomes, but affecting student performance while still in high school.

Finally, we analyze the impact of initial failure on the CAHSEE on post-secondary outcomes. The National Student Clearinghouse data for students in SDUSD provides us with the ability to investigate whether initial failure on the exit exam affects student’s long term academic career. We investigate the probability of enrolling in 2-year and 4-year institutions. Because transferring from a community college to a 4-year college is a common strategy for students, we also investigate whether initial failure on the exam affects the probability that a student transfers between 2- and 4-year colleges. The analysis of post-secondary outcomes sheds light on whether a failure on the CAHSEE could have long term negative consequences for students and separates our analysis from the literature which focused on high school outcomes only.

We first analyze students who fail the CAHSEE Math in grade 10 and investigate whether passing the CAHSEE ELA test causes any discernible change in the
outcomes of interest. Second, we carry out the analysis in the sample of students who passed the CAHSEE Math test in grade 10. We continue the analysis by dividing the sample along the ELA passing cutoff: we analyze the effect of passing the CAHSEE Math test in grade 10 among students who failed the CAHSEE ELA test in grade 10 and among the students who passed the CAHSEE ELA exam in grade 10, separately. In each case, the regression estimates presented summarize the results for the analysis where we control for student characteristics in grade nine, school level time varying characteristics in grade eleven for outcomes in grade eleven and school varying characteristics in grade eleven and twelve for the graduation probability and post-secondary outcomes. In addition all models include school fixed effects and time fixed effects. Tables of results contain four panels, where each panel summarizes results for the separate student sub-population. We present four sets of estimates for each sub-population to account for the four bandwidths discussed in the bandwidth selection section.

2.5.1 Remedial Courses and Academic Trajectories

The theoretical framework presented in Section 3 predicts that failure on either component of the CAHSEE should cause a discrete jump in the probability of the student receiving remedial instruction. We would expect that the students who failed the math component would have a higher probability of enrolling in a math prep course, while those who failed ELA would have a higher probability associated with enrollment in an ELA prep course. Furthermore, we would expect that failure in math should not affect the probability of enrollment in ELA and vice versa.

The estimation results confirm the above conjecture; therefore we present only the results relevant to enrollment in ELA and math prep courses for the population for which the failure on the CAHSEE should affect remedial course taking. Table 2.3
presents the regression results for model of the effects of CAHSEE failure in grade 10 on the probability of enrollment in ELA remedial courses in term 1 and term 2 of grade 11. The results confirm that initial failure on the CAHSEE increases the probability of enrollment in preparatory courses among students close to the passing cutoff, as predicted. In all four panels the estimated coefficients are not statistically significant for the smallest bandwidth used; however, with the increase in bandwidth the effects of failing CAHSEE ELA in grade 10 become statistically significant.

Among the students who failed CAHSEE math (top left panels), those who failed CAHSEE ELA have between 4.5 and 5.7 percentage points higher probability of enrolling in a CAHSEE ELA preparatory course in the first term of grade 11, as compared to students who failed Math but passed ELA in grade 10. Failure in ELA in grade 10 is also associated with an increase in the likelihood of enrollment in a preparatory class in the second term in grade 11, but the results are not as strong as for term 1.

TABLE 2.3 HERE

Among students who passed math (top right panel), those who failed ELA are 6.4 to 6.5 percentage points more likely to enroll in an ELA prep class in the first term of grade 11 than those who passed ELA. Again, the effect estimated using bandwidth of 10 points are statistically not significant. Similar pattern emerges for term 2. For a student who failed CAHSEE math in grade 10 (bottom left panel) the likelihood of enrolling in an ELA preparatory class in the second term of grade 11 increases by 1.8 to 2.5 percentage points if the student fails CAHSEE ELA in grade 10. However, the estimated effects are statistically significant at 5% level only with bandwidth of 30 and 40 points around the passing cutoff. The estimated effects are not surprising,
as the students who struggled with the test on their first attempt should also be the ones seeking additional help before they have to retake the exam.

Graphically, the results are presented in Figure 2.1. Due to the fact that students do not have to retake parts of the exam they have passed, the probability of enrollment in a CAHSEE prep class for students who pass a given component should be zero in all cases. The deviation from that behavior can be explained by students enrolling in prep classes as electives- given that they passed the CAHSEE test, the class should be an easy one for them and might satisfy the elective requirements. The data to the left of the cutoff in each panel displays a larger degree of variation. Overall, the graphical analysis confirms the regression results from Table 2.3.

FIGURE 2.1 HERE

The pattern of effects of failure on CAHSEE Math and the probability of enrollment in math preparatory classes in grade 11 is similar to the effects presented for the ELA prep classes. The effect of failing CAHSEE Math in grade 10 within 10 points of the passing cutoff among students who also failed ELA (top left panel) has a negative sign, but is statistically not significant. With the increase in bandwidth, the effect changes sign and become statistically significant at the 1% level. The results suggest that at the margin the students behave very similarly, but as we increase the bandwidth, students who initially failed are more likely to enroll in preparatory classes in term 1 of grade 11. Among students who passed CAHSEE ELA in grade 10 (top right panel), failure on CAHSEE Math is strongly associated with an increase in enrollment in Math prep class in first term of grade 11. The estimated coefficient ranges from 2.9 to 5.9 percentage points and is statistically significant at 1% level across all specified bandwidths. The effects of math failure in
grade 10 on the likelihood of enrollment in math prep classes in term 2 of grade 11 (bottom panels) are similar to the effects on enrollment in term 1. Students who also failed (passed) ELA are at most 2.9 (2.1) percentage points more likely to enroll in a math prep course in term 2 of grade 11.

2.5.2 Test Failure and Performance in High School

The CAHSEE content raises concerns whether the test is pitched too low to change student motivation and long-term achievement (Zau and Betts (2008)). However, a performance related signal might also have short-term effects. Thus, we would expect that a signal acquired from the CAHSEE could induce a change in performance the following year. Consider first the student score on the California Standards Test (CST) in English Language Arts and in math. The CST is aligned with grade and course specific standards and therefore it is relevant to material covered in classes taken by the student. If failure on the CAHSEE in grade 10 increases student motivation or provides students with remedial instruction to assure eventual success on the exit exam, it could also translate into better performance on the CST. The effects associated with achievement on the CST rest on the assumption that the material required for success on the CAHSEE is also within the test domain of the CST. If the material tested on the grade 11 CST has no overlap with the knowledge domain associated with earlier grades, the effects of the failure and remedial help could be limited to improved test taking skills. Under this scenario failure on the exit exam in grade 10 could have no effect on student performance on the grade 11 CST.

Grade point average (GPA) might provide us with insights into whether the initial failure on the exam changes student effort and achievement. The rationale for looking at GPA is the fact that grades have a subjective component associated with the fact that teachers take into consideration student’s performance, motivation, par-
ticipation and overall classroom behavior. Grades might capture aspects of student behavior not detectable through test scores. Increased attention by teachers towards the students who barely failed the CAHSEE in the previous year, potentially combined with a change in student effort, might lead to a change in performance in the classroom. As a result the initial failure on the CAHSEE might induce a significant change in GPA the following year.

Table 2.4 reports the results of the effects of initial failure on the CAHSEE in grade 10 on student achievement on the CST ELA, CST Math, and GPA in grade 11. We report regression results using the 20 point bandwidth for each of the three outcomes discussed above.

The results shown in Table 2.4 suggest that failing the CAHSEE in grade 10 has no effect on student performance in grade 11, as measured by the low-stakes CST scores and GPA. Not only are the coefficients statistically insignificant but they are small in magnitude. The results are further shown graphically in Figures 2.3 and 2.4. The results presented so far suggest that the initial failure induces students just below the passing cutoff to enroll in remedial courses, but the enrollment in those courses does not seem to affect student performance on standardized tests and does not seem to have an effect on student grades in the following year. The empirical results suggest that the remedial investments made after the first attempt at the CAHSEE do not affect the overall pattern of achievement. Given the results, the theoretical framework suggest that the RD results should capture the composite effects of the initial failure.

\footnote{The remaining estimation results are included in the Appendix as in all four cases the estimated effects are similar across the bandwidths used in the analysis. Table A.2.2 summarizes the effects for the low-stakes CST ELA scores, Table A.2.3 the effects for the low-stakes CST Math scores, and Table A.2.4 for GPA.}
Early failure on the exit exam seems to have very little effect on student achievement while in high school among students close to the passing cutoff. We turn our attention towards the potential effects of the CAHSEE on the probability of dropping out of high school and the probability of graduating with a diploma. Failure on the CAHSEE on the first attempt, could, in theory, translate into failure to graduate from high school. Initial performance on the exam carries direct significance for whether or not the student will have to take the test again and therefore, it might significantly affect the probability of graduating from high school with a diploma by the end of grade twelve. It might also increase the likelihood that a student drops out.

In Table 2.5 we present the results of the effect of failing CAHSEE in grade 10 on the probability of dropping out after grade 10, the probability of graduating from high school within 4 years of first enrolling in grade 9, and the probability of graduating from high school within 6 years of first enrolling in grade 9. As in previous sections, the results are estimated within 10 points from the passing cutoff. Additional specification with varying bandwidths for each of the outcomes are provided in the Appendix.
We define dropout as a student who was confirmed by the school district to have dropped out after grade 10\(^6\). Table 2.5 shows that for all student populations considered in the analysis, initial failure on the CAHSEE has no effect on the likelihood of dropping out after grade 10 among students close to the passing cutoff. Furthermore, being just below the passing cutoff on either of the components of the CAHSEE does not affect the likelihood of graduation from high school, whether we consider on-time graduation (within four years of enrolling in grade 9 for the first time), or the probability of graduation within six years of enrolling in grade 9 for the first time. For students just below the passing cutoff the initial failure on the exit exam does not cause graduation delays, further confirming that the exam is not a real barrier for the students just below the cutoffs.

### 2.5.4 Postsecondary Outcomes

Poor initial performance on the exit exam does not affect student achievement while in high school. However, it is possible that it could affect students’ choices associated with postsecondary education. The additional practice in test taking techniques associated with the preparatory courses and any remedial work done by the student as well as any other help from teachers in order to assure success on the CAHSEE could translate into better preparedness for test taking in the future. Thus, it is possible that students who initially fail the CAHSEE might have different patterns of postsecondary enrollment than students who did not have to retake the exit exam. Given that the students just above and just below the passing cutoff in grade 10 have similar probabilities of graduating from high school, they are in the same position to decide to enroll in postsecondary institutions. Hence any systematic differences in

---

\(^6\)The definition of a dropout is a lot more conservative than in many other studies, as we want to make sure that the observations we count as dropouts are not students who transferred districts. That way we avoid overstates of the number of dropouts.
enrollment decisions could be associated with the initial failure on the CAHSEE.

We model three distinct outcomes: the probability of enrollment in a two-year college, the probability of enrollment in a four-year university (both within two years on finishing high school), and the likelihood that a student initially enrolled in a two-year college ever transfers to a four-year institution, conditional on enrollment in a 2-year college. In each case we report the estimates for the regressions using the 20 point bandwidth and report the remaining results for each of the outcomes in the Appendix. Table 2.6 summarizes the results of the estimation of the effects of CAHSEE failure on the probability of postsecondary enrollment among students close to the passing cutoff as well as the likelihood of transfer to a four-year college.

TABLE 2.6 HERE

Similar to previous sections, the regressions suggest no effect of initial failure on the CAHSEE on the probability of enrollment in college, or transfer between 2- and 4-year college. The only notable exception is the probability of enrolling in a four-year college among students who failed CAHSEE Math in grade 10. For the students who failed CAHSEE Math in grade 10, failing CAHSEE ELA is associated with a 4.4 percentage points higher probability of enrolling in a four-year college than the students who barely passed ELA in grade 10. The estimated effect is statistically significant at the 5% level. Additional results associated with the probability of enrollment in a 4-year program are presented in Table A.2.9. The increase in bandwidth decreases the magnitude of the effect. However, even for the bandwidth of 40 points, the students who fail CAHSEE ELA in grade 10 have a 2.9 percentage point higher probability of enrollment in a 4-year college, statistically significant at the 5% level. In contrast, if the bandwidth is reduced to 10 points around the passing
cutoff, the magnitude of the effect remains stable, but the coefficient loses statistical significance.

The interpretation of the result associated with failure in ELA among the students who also failed CAHSEE Math in grade 10 is challenging, as the results presented in the analysis do not point towards a difference in high-school level outcomes. The only notable difference between the students just below and just above the cutoff is the higher probability of enrolling in a CAHSEE preparatory classes in grade 11 for the students who failed ELA in grade 10. If the prep courses provided students just below the cutoff with test-taking skills that improve their ability to successfully pass high-stakes exams, such as the SAT and perhaps college exams, it could be considered as potentially impacting long term academic trajectories of these students. However, if this argument were true, then we would expect a similar effect among the students who passed CAHSEE Math in grade 10. Those students also have a higher probability of enrolling in CAHSEE ELA prep courses, but their estimated coefficient of failure on the probability of enrollment in a 4-year college is statistically insignificant.

2.6 Conclusions

The empirical analysis guided by the proposed theoretical framework suggests very limited effects of initial failure on the CAHSEE in grade 10 on student achievement and postsecondary outcomes among students from San Diego Unified School District close to the passing cutoff in either math or ELA. We find that students who fail a portion of the exam are significantly more likely to enroll in preparatory courses designed specifically for success on the CAHSEE. We show that initial failure on the exam has no effect on student achievement while in high school. The analysis of the post-secondary outcomes reveals similar pattern to the one found for the high school outcomes. Scoring just below the passing cutoff in math or ELA does not
significantly affect most postsecondary outcomes. The notable difference is the effect among students who failed CAHSEE Math. In that student subgroup the failure on CAHSEE ELA is associated with a higher likelihood of enrollment in a four-year college. Unfortunately we don’t have an insight into the mechanism through which the effect might be operating and statements regarding positive effects should be made with caution.

The regression results suggest that initial failure on the exit exam does not have negative long term effects for student outcomes. The analysis presented is consistent with the literature on exit examinations. It is unclear whether the estimated effects result from increased student effort just below the cutoff and increased teacher attention, or possibly a decrease in effort among the students just above the minimum passing scores. However, the increased likelihood in remedial course enrollment points towards the former. The empirical findings can also be interpreted as a result of an introduction of a uniform standard pitched at a low level of required competency, which induces no change in student achievement, because the standard is low.

The fact that both running variables that determine the discontinuities are discrete implies that the conventional error structure might pose a problem in our context. The analysis presented rests on the assumption that the error structure is the same regardless of which side of the cutoff we consider. In reality the error structures above and below the cutoff might be different. Typically, this recognition would motivate the use of the Lee and Card (2008) error correction. However in the context where all of the estimated results are statistically insignificant, the error correction, which tends to increase the size of the error terms, seems unnecessary.

The results presented in this paper strongly suggest the lack of an effect of initial failure on the CAHSEE on any of the major academic outcomes. The empirical results in conjunction with the theoretical framework discussed suggest that a test
such as the CAHSEE has the scope to provide students with important information regarding their position relative to the standard. However, the data available does not allow us to further unpack the potential mechanisms generating the estimated zero effects. We recognize that the regression discontinuity framework results are local average treatment effects within a narrow window of the cutoff score. We confirm the findings of Reardon et al. (2010) and further show that the exam has no effect on postsecondary outcomes. The analysis suggest that CAHSEE is pitched at a level too low to cause distortions to academic trajectories of students close to the passing cutoff on the initial attempt at the exam.

The lack of negative effects associated with early failure on the CAHSEE are reassuring. The analysis shows that the exam does not cause negative distortions in the educational system, at least among the students in the close neighborhood of the passing cutoff. Students who initially fail the exam by a small margin are more likely to seek remedial help, as compared to the students who barely passed, but they do not fall behind in terms of their academic achievement. In the context of a regression discontinuity analysis absence of statistically significant negative effects should be viewed as a positive feature of the exit exam system put in place in California.
### 2.7 Tables

**Table 2.1:** Means of Baseline Student Characteristics in Grade 9 by Grade 10 CAHSEE Performance Subgroup

<table>
<thead>
<tr>
<th></th>
<th>Failed Math</th>
<th>Failed Math</th>
<th>Passed Math</th>
<th>Passed Math</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Failed ELA</td>
<td>Passed ELA</td>
<td>Failed ELA</td>
<td>Passed ELA</td>
</tr>
<tr>
<td>CST Math</td>
<td>-0.518</td>
<td>-0.547</td>
<td>-0.421</td>
<td>-0.464</td>
</tr>
<tr>
<td>CST ELA</td>
<td>-0.555</td>
<td>-0.354</td>
<td>-0.480</td>
<td>-0.351</td>
</tr>
<tr>
<td>GPA</td>
<td>2.046</td>
<td>2.001</td>
<td>2.043</td>
<td>2.094</td>
</tr>
<tr>
<td>% Days Absent</td>
<td>4.754</td>
<td>4.941</td>
<td>4.366</td>
<td>4.439</td>
</tr>
<tr>
<td>English Learner</td>
<td>0.428</td>
<td>0.320</td>
<td>0.408</td>
<td>0.275</td>
</tr>
<tr>
<td>Special Education</td>
<td>0.143</td>
<td>0.141</td>
<td>0.143</td>
<td>0.141</td>
</tr>
<tr>
<td>Female</td>
<td>0.472</td>
<td>0.554</td>
<td>0.479</td>
<td>0.498</td>
</tr>
<tr>
<td>African-American</td>
<td>0.174</td>
<td>0.205</td>
<td>0.152</td>
<td>0.183</td>
</tr>
<tr>
<td>Asian</td>
<td>0.129</td>
<td>0.106</td>
<td>0.111</td>
<td>0.121</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.612</td>
<td>0.577</td>
<td>0.643</td>
<td>0.582</td>
</tr>
<tr>
<td>Observations</td>
<td>621</td>
<td>672</td>
<td>532</td>
<td>759</td>
</tr>
</tbody>
</table>

Note: table of means within each subgroup based on CAHSEE Math and ELA scores in grade 10. The means presented are provided for a window within 10 points of the passing cutoff score.

**Table 2.2:** Tests of Equality of Means of Grade 9 Characteristics by Grade 10 CAHSEE Performance Subgroup

<table>
<thead>
<tr>
<th></th>
<th>Failed Math</th>
<th>Passed Math</th>
<th>Failed ELA</th>
<th>Passed ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Failed vs. Passed ELA</td>
<td>Failed vs. Passed ELA</td>
<td>Failed vs. Passed Math</td>
<td>Failed vs. Passed Math</td>
</tr>
<tr>
<td>CST Math</td>
<td>-0.187</td>
<td>1.141</td>
<td>2.354**</td>
<td>4.364**</td>
</tr>
<tr>
<td>CST ELA</td>
<td>5.519**</td>
<td>5.126**</td>
<td>-0.152</td>
<td>1.243</td>
</tr>
<tr>
<td>GPA</td>
<td>1.658*</td>
<td>0.662</td>
<td>1.340</td>
<td>1.846*</td>
</tr>
<tr>
<td>% Days Absent</td>
<td>0.167</td>
<td>-1.563</td>
<td>-1.340</td>
<td>-2.968**</td>
</tr>
<tr>
<td>English Learner</td>
<td>3.151</td>
<td>14.836**</td>
<td>0.051</td>
<td>1.567</td>
</tr>
<tr>
<td>Special Education</td>
<td>0.118</td>
<td>0.936</td>
<td>1.477</td>
<td>0.635</td>
</tr>
<tr>
<td>Female</td>
<td>1.513</td>
<td>0.195</td>
<td>0.004</td>
<td>1.891</td>
</tr>
<tr>
<td>African-American</td>
<td>0.055</td>
<td>2.129</td>
<td>0.008</td>
<td>0.761</td>
</tr>
<tr>
<td>Asian</td>
<td>1.145</td>
<td>0.893</td>
<td>0.611</td>
<td>1.022</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.088</td>
<td>9.712**</td>
<td>0.000</td>
<td>0.296</td>
</tr>
</tbody>
</table>

Note: continuous variables - t-test of equality of means; categorical variables - Pearson Chi-squared test. The tests carried out within 10 points of the passing cutoff score. ** p < 0.01; * p < 0.05.
Table 2.3: Effects of Failing CAHSEE in Grade 10 on the Probability of Enrolling in a CAHSEE ELA Preparatory Class in Grade 11

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.027</td>
<td>0.057*</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,438</td>
<td>2,903</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.156</td>
<td>0.105</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level
** p < 0.01, * p < 0.05

Table 2.4: Effects of Failing CAHSEE in Grade 10 on Student Achievement on ELA Low-Stakes Tests in Grade 11

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Failed ELA in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ELA Score</td>
<td>Math Score</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.024</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Observations</td>
<td>3,410</td>
<td>2,903</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.133</td>
<td>0.294</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Failed Math in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ELA Score</td>
<td>Math Score</td>
</tr>
<tr>
<td>Failed Math</td>
<td>0.029</td>
<td>-0.030</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.044)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,596</td>
<td>2,171</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.145</td>
<td>0.279</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level
** p < 0.01, * p < 0.05
### Table 2.5: Effects of Failing CAHSEE in Grade 10 on High School Completion Outcomes

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grad Dropout On-Time in 6-Yr</td>
<td>Grad Dropout On-Time in 6-Yr</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>-0.013, 0.027, 0.099</td>
<td>-0.013, 0.039, 0.024</td>
</tr>
<tr>
<td>Observations</td>
<td>4,517, 2,372, 1,667</td>
<td>5,117, 2,542, 1,557</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.075, 0.283, 0.230</td>
<td>0.056, 0.288, 0.252</td>
</tr>
</tbody>
</table>

Failed ELA in Gr10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Failed ELA in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grad Dropout On-Time in 6-Yr</td>
<td>Grad Dropout On-Time in 6-Yr</td>
</tr>
<tr>
<td>Failed Math</td>
<td>0.018, -0.020, -0.073</td>
<td>0.009, -0.038, -0.042</td>
</tr>
<tr>
<td>Observations</td>
<td>3,594, 1,735, 442</td>
<td>7,195, 4,287, 3,006</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.067, 0.278, 0.258</td>
<td>0.045, 0.296, 0.238</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05

### Table 2.6: Effects of Failing CAHSEE in Grade 10 on Post-Secondary Academic Outcomes

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>College 2 College 4 Transfer</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>-0.001, 0.044* -0.006</td>
<td>-0.019, 0.020 -0.026</td>
</tr>
<tr>
<td>Observations</td>
<td>1,995, 1,995 633</td>
<td>2,003, 2,003 862</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.164, 0.070 0.139</td>
<td>0.123, 0.070 0.128</td>
</tr>
</tbody>
</table>

Failed ELA in Gr10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Failed ELA in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>College 2 College 4 Transfer</td>
</tr>
<tr>
<td>Failed Math</td>
<td>0.054, 0.005 0.015</td>
<td>0.029, 0.004 0.027</td>
</tr>
<tr>
<td>Observations</td>
<td>1,501, 1,501 442</td>
<td>3,301, 3,301 1,496</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.137, 0.069 0.258</td>
<td>0.088, 0.095 0.131</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05
2.8 Figures

![Figures](image1)

**Note:** top - effect of failing the CAHSEE ELA on probability of enrollment in an ELA prep class among students who failed CAHSEE Math in grade 10; bottom - effect of failing the CAHSEE ELA on probability of enrollment in an ELA prep class among students who passed CAHSEE Math in grade 10.

**Figure 2.1:** The Effects of Failing CAHSEE ELA in Grade 10 on the Probability of Enrollment in an ELA Preparatory Course in Grade 11

![Figures](image2)

**Note:** top - effect of failing the CAHSEE Math on probability of enrollment in an Math prep class among students who failed CAHSEE ELA in grade 10; bottom - effect of failing the CAHSEE Math on probability of enrollment in a Math prep class among students who passed CAHSEE ELA in grade 10.

**Figure 2.2:** The Effects of Failing CAHSEE Math in Grade 10 on the Probability of Enrollment in an Math Preparatory Course in Grade 11
**Figure 2.3:** The Effects of Failing CAHSEE ELA in Grade 10 on Student Outcomes in Grade 11

**Figure 2.4:** The Effects of Failing CAHSEE Math in Grade 10 on Student Outcomes in Grade 11
2.9 Validity of the RD Framework

In order to claim that the RD framework is appropriate in our context we have to make sure that our data satisfy two main conditions. Firstly, the forcing variables, the CAHSEE Math and the CAHSEE ELA scores, have to be smooth around the cutoff. A lack of smoothness or a visible discontinuity in the forcing variable would suggest that the cutoff could be manipulated by some individuals. In such a case the assertion that the individuals just above and below the cutoff are comparable could not be sustained. The students who could manipulate their score to be just above the cutoff would be significantly different than the students who could not manipulate their score in the same way and ended up below the cutoff.

In our case we have to worry not only about the overall distribution of CAHSEE scores in each subject, but also about the combination of the CAHSEE scores and their joint distribution. If we consider each score separately it may appear that we meet the smoothness condition. At the same time it is possible, that for particular subgroups, for example the students who passed CAHSEE Math, the distributions of the CAHSE ELA scores are not smooth around the cutoff. In such a case the smoothness condition would be violated.

We show that for the students who failed CAHSEE Math in grade 10 the distribution of the CAHSEE ELA scores is smooth, the same for students who passed CAHSEE Math. Additionally the data are smooth around the Math cutoff both for students who failed CAHSEE ELA in grade 10 as well as for those who passed CAHSEE ELA. We present histograms of data only around the cutoffs as those are the relevant areas in which we need to make sure that the data satisfies the smoothness conditions.

Figure A.2.1 shows the histograms of the scaled scores in each subject in each sub-population under investigation within the twenty point window around the cutoff.
score. In all cases our data satisfies the smoothness condition. There does not seem to be any manipulation around the cutoffs.

FIGURE A.2.1 HERE

We further test this by collapsing the data at the score level and testing whether more or less students are reported just below the cutoff or just above the cutoff (Almond, Doyle Jr., Kowalski, and Williams (2010)). We run the following regression:

\[ y_{ij} = \beta_0 + \beta_1 \times 1[CAHSEE_i < 0] + \beta_2 CAHSEE_i \times 1[CAHSEE_i < 0] + \beta_3 CAHSEE_i \times 1[CAHSEE_i \geq 0] + \beta_4 CAHSEE_{-i} + \varepsilon_{ij} \] (2.5)

where \( y_{ij} \) is the number of observations of \( CAHSEE_i \), \( i \in \{\text{Math, ELA}\} \) for a given standardized \( CAHSEE \) score, \( 1[CAHSEE_i < 0] \) is an indicator function equal to one if the standardized \( CAHSEE_i \) score is below the passing cutoff, zero otherwise; \( 1[CAHSEE_i \geq 0] \) is an binary variable equal to one if the standardized \( CAHSEE_i \) score is equal to or above the passing cutoff, zero otherwise and \( CAHSEE_{-i} \) is the standardized score in the other subject tested. The regression results are presented in Table A.2.11.

TABLE A.2.11 HERE

In the sample of students who failed CAHSEE Math and were within the twenty point window around the ELA cutoff, the coefficient on the cutoff dummy is 5.868 with a robust standard error of 47.275. For students who passed CAHSEE Math the coefficient is -16.755 with a robust standard error of 57.52. Among the
students who failed CAHSEE ELA in grade ten the coefficient on the failed math dummy is equal to -17.15 with a robust standard error of 46.385. Finally for students who passed CAHSEE ELA in grade ten the coefficient is equal to -60.841 with a robust standard error of 89.85. These test results further confirm that there is no manipulation of the running variable in either of the cases under analysis.

Finally we investigate the potential for the discontinuity of the control variables around the cutoff using Equation 2.5. Given that the regression discontinuity assumptions supposed to hold at the limit, we choose a bandwidth of four points around the passing cutoff to compare students who should be most similar. The regression results are presented in Table ???. The control variables pertaining to student performance appear to be continuous around the passing score cutoffs in majority of cases. For 3 of the 36 regressions the result suggest significant differences below and above the cutoffs.

The data suggest that a smaller fraction of English Learners and a larger fraction of females fall just below the passing cutoff in math, among the students who passed CAHSEE ELA in grade 10. Additionally, the students just below the passing cutoff in the same sub-sample are absent 1.85 percentage points of the school year more than those just above the cutoff. The size of the effect on absences is very small and to large extent can be ignored. In order to address the differences in terms of student composition below and above the cutoffs we control for the characteristics of the student population used in the analysis. The data seems to be satisfying the conditions required for identification in the regression discontinuity framework. The overall pattern in all four sub-populations looks similar, confirming that the data are well suited for a regression discontinuity analysis.

TABLE A.2.12 HERE

TABLE A.2.13 HERE
2.10 Supplementary Tables and Figures

![Histogram Plots](image)

Note: top left - ELA Scores in gr10, students who failed CAHSEE Math in gr10, top right - ELA Scores in gr10, students who passed CAHSEE Math in gr10; bottom left - Math Scores in gr10, students who failed CAHSEE ELA in gr10; bottom right - Math Scores in gr10, students who passed CAHSEE ELA in gr10.

**Figure A.2.1:** Histogram Plots of the Scaled CAHSEE ELA and Scaled CAHSEE Math Scores in Grade 10

**Table A.2.1:** Effects of Failing CAHSEE in Grade 10 on the Probability of Enrolling in CAHSEE Math Preparatory Class in Grade 11

<table>
<thead>
<tr>
<th></th>
<th>Failed ELA in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td>bandwidth:</td>
<td>10 20 30 40</td>
<td>10 20 30 40</td>
</tr>
<tr>
<td>Failed Math</td>
<td>-0.015 0.033** 0.024* 0.020*</td>
<td>0.055** 0.038** 0.031** 0.029**</td>
</tr>
<tr>
<td></td>
<td>(0.018) (0.013) (0.011) (0.010)</td>
<td>(0.015) (0.009) (0.008) (0.007)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,063 2,171 3,178 4,216</td>
<td>2,534 5,087 7,473 9,961</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.154 0.114 0.103 0.093</td>
<td>0.118 0.087 0.073 0.071</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Failed ELA in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td>bandwidth:</td>
<td>10 20 30 40</td>
<td>10 20 30 40</td>
</tr>
<tr>
<td>Failed Math</td>
<td>0.010 0.029* 0.018* 0.021*</td>
<td>0.021* 0.011 0.011 0.016**</td>
</tr>
<tr>
<td></td>
<td>(0.018) (0.012) (0.011) (0.009)</td>
<td>(0.009) (0.008) (0.006) (0.006)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,063 2,171 3,178 4,216</td>
<td>2,534 5,087 7,473 9,961</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.082 0.073 0.065 0.063</td>
<td>0.067 0.077 0.07 0.074</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05
Table A.2.2: Effects of Failing CAHSEE in Grade 10 on Student Achievement on ELA Low-Stakes Tests in Grade 11

<table>
<thead>
<tr>
<th></th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 20 30 40</td>
<td>10 20 30 40</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.065 0.024 0.006 0.012</td>
<td>0.065 0.020 -0.013 0.009</td>
</tr>
<tr>
<td></td>
<td>(0.058) (0.041) (0.033) (0.029)</td>
<td>(0.057) (0.039) (0.032) (0.028)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,684 3,410 4,989 6,417</td>
<td>1,809 3,701 5,587 7,923</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.139 0.133 0.141 0.162</td>
<td>0.167 0.189 0.209 0.249</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

**p < 0.01, *p < 0.05

Table A.2.3: Effects of Failing CAHSEE in Grade 10 on Student Achievement on Math Low-Stakes Tests in Grade 11

<table>
<thead>
<tr>
<th></th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 20 30 40</td>
<td>10 20 30 40</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.027 0.029 0.023 0.028</td>
<td>-0.056 -0.060 -0.009 -0.006</td>
</tr>
<tr>
<td></td>
<td>(0.072) (0.046) (0.038) (0.033)</td>
<td>(0.047) (0.032) (0.027) (0.024)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,274 2,596 3,813 5,061</td>
<td>2,818 5,656 8,284 11,014</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.165 0.145 0.130 0.130</td>
<td>0.331 0.326 0.332 0.353</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

**p < 0.01, *p < 0.05
Table A.2.4: Effects of Failing CAHSEE in Grade 10 on Grade Point Average in Grade 11

<table>
<thead>
<tr>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td>bandwidth:</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,438</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.772</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Failed ELA in Gr10</th>
<th>Passed ELA in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td>bandwidth:</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed Math</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>(0.078)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,063</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.841</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05

Table A.2.5: Difference-in-Difference Estimate of the Impact of the CAHSEE on the Probability of Dropping Out by Decile of Initial Achievement (dp/dx)

<table>
<thead>
<tr>
<th>Decile</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAHSEE 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>-0.005</td>
<td>0.001</td>
<td>-0.005</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.006)</td>
<td>(0.005)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>CAHSEE 2</td>
<td>0.011**</td>
<td>0.004</td>
<td>0.010**</td>
<td>-0.005</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.003)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>Decile</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
<td>(9)</td>
<td>(10)</td>
</tr>
<tr>
<td>CAHSEE 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.011**</td>
<td>-0.003</td>
<td>0.000</td>
<td>-0.005</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.007)</td>
<td>(0.012)</td>
<td>(0.005)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>CAHSEE 2</td>
<td>-0.008**</td>
<td>0.001</td>
<td>0.002</td>
<td>-0.011**</td>
<td>-0.012***</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.005)</td>
<td>(0.004)</td>
</tr>
</tbody>
</table>

Full Set of Controls ✓

Pseudo R-squared: 0.209 N: 48,000 DVM: 0.037

Standard errors in parentheses, clustered at the school level.
** p<0.01; * p<0.05
Table A.2.6: Effects of Failing CAHSEE in Grade 10 on the Probability of Graduating from High School 4 Years after Enrolling in Grade 9

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.064</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>(0.055)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,176</td>
<td>2,372</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.396</td>
<td>0.283</td>
</tr>
</tbody>
</table>

Failed ELA in Gr10

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math</th>
<th>Passed ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed Math</td>
<td>0.015</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td>(0.044)</td>
</tr>
<tr>
<td>Observations</td>
<td>825</td>
<td>1,735</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.425</td>
<td>0.278</td>
</tr>
</tbody>
</table>

Failed ELA in Gr10

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math</th>
<th>Passed ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed Math</td>
<td>-0.067</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>(0.088)</td>
<td>(0.063)</td>
</tr>
<tr>
<td>Observations</td>
<td>502</td>
<td>1,100</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.504</td>
<td>0.249</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level
** p< 0.01, * p< 0.05

Table A.2.7: Effects of Failing CAHSEE in Grade 10 on the Probability of Graduating from High School Within 6 Years of Enrolling in Grade 9

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.131</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>Observations</td>
<td>793</td>
<td>1,667</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.394</td>
<td>0.230</td>
</tr>
</tbody>
</table>

Failed ELA in Gr10

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math</th>
<th>Passed ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed Math</td>
<td>-0.067</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>(0.088)</td>
<td>(0.063)</td>
</tr>
<tr>
<td>Observations</td>
<td>502</td>
<td>1,100</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.504</td>
<td>0.249</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level
** p< 0.01, * p< 0.05
### Table A.2.8: The Effect of Failing CAHSEE in Grade 10 on the Probability of Enrolling in a 2-Year College After High School

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.070</td>
<td>0.002</td>
</tr>
<tr>
<td>20</td>
<td>-0.001</td>
<td>-0.019</td>
</tr>
<tr>
<td>30</td>
<td>0.002</td>
<td>-0.041</td>
</tr>
<tr>
<td>40</td>
<td>0.033</td>
<td>-0.034</td>
</tr>
<tr>
<td>10</td>
<td>(0.068)</td>
<td>(0.071)</td>
</tr>
<tr>
<td>20</td>
<td>(0.045)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>30</td>
<td>(0.037)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>40</td>
<td>(0.032)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.002</td>
<td>0.011</td>
</tr>
<tr>
<td>10</td>
<td>0.021</td>
<td>0.009</td>
</tr>
<tr>
<td>20</td>
<td>0.044</td>
<td>0.013</td>
</tr>
<tr>
<td>30</td>
<td>(0.035)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>40</td>
<td>(0.019)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>Observations</td>
<td>989</td>
<td>1,030</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.164</td>
<td>0.174</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05

### Table A.2.9: The Effect of Failing CAHSEE in Grade 10 on the Probability of Enrolling in a 4-Year College After High School

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.041</td>
<td>0.013</td>
</tr>
<tr>
<td>20</td>
<td>0.044</td>
<td>0.020</td>
</tr>
<tr>
<td>30</td>
<td>0.038</td>
<td>0.006</td>
</tr>
<tr>
<td>40</td>
<td>0.029</td>
<td>0.005</td>
</tr>
<tr>
<td>10</td>
<td>(0.035)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>20</td>
<td>(0.020)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>30</td>
<td>(0.015)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>40</td>
<td>(0.013)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>0.041</td>
<td>0.013</td>
</tr>
<tr>
<td>10</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td>20</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>30</td>
<td>0.017</td>
<td>0.003</td>
</tr>
<tr>
<td>40</td>
<td>0.023</td>
<td>-0.005</td>
</tr>
<tr>
<td>10</td>
<td>(0.044)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>20</td>
<td>(0.020)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>30</td>
<td>(0.014)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>40</td>
<td>(0.013)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Observations</td>
<td>989</td>
<td>1,030</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.102</td>
<td>0.091</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05
**Table A.2.10:**  The Effect of Failing CAHSEE in Grade 10 on the Probability of Transferring From a 2-Year to a 4-Year College

<table>
<thead>
<tr>
<th>bandwidth:</th>
<th>Failed Math in Gr10</th>
<th>Passed Math in Gr10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>-0.047</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(0.060)</td>
</tr>
<tr>
<td>Observations</td>
<td>310</td>
<td>633</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.298</td>
<td>0.139</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

**Table A.2.11:**  Robustness Checks: Manipulation of the Running Variable

<table>
<thead>
<tr>
<th>sub-sample:</th>
<th>Failed Passed Failed Passed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Math in Gr10 Math in Gr10 ELA in Gr10 ELA in Gr10</td>
</tr>
<tr>
<td>Failed ELA</td>
<td>5.868 -16.755</td>
</tr>
<tr>
<td></td>
<td>(47.275) (57.520)</td>
</tr>
<tr>
<td>Failed Math</td>
<td>-17.150 -60.841</td>
</tr>
<tr>
<td></td>
<td>(46.385) (89.850)</td>
</tr>
</tbody>
</table>

Observations collapsed at the score level to test for manipulation of the forcing variable.

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

**Notes:**

- **p < 0.01, * p < 0.05**
Table A.2.12: Robustness Checks: Testing for Breaks in the Achievement Control Variables in Grade 10

Failed CAHSEE Math in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>CST ELA</th>
<th>CST Math</th>
<th>GPA</th>
<th>% Absent</th>
<th>Learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed ELA</td>
<td>-0.092</td>
<td>0.115</td>
<td>-0.168</td>
<td>0.922</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>(0.073)</td>
<td>(0.078)</td>
<td>(0.116)</td>
<td>(0.888)</td>
<td>(0.075)</td>
</tr>
<tr>
<td>Observations</td>
<td>798</td>
<td>722</td>
<td>792</td>
<td>798</td>
<td>798</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.014</td>
<td>0.022</td>
<td>0.030</td>
<td>0.004</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Passed CAHSEE Math in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>CST ELA</th>
<th>CST Math</th>
<th>GPA</th>
<th>% Absent</th>
<th>Learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed ELA</td>
<td>-0.032</td>
<td>-0.081</td>
<td>-0.188</td>
<td>1.743</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>(0.119)</td>
<td>(0.173)</td>
<td>(0.106)</td>
<td>(1.205)</td>
<td>(0.087)</td>
</tr>
<tr>
<td>Observations</td>
<td>698</td>
<td>665</td>
<td>693</td>
<td>698</td>
<td>698</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.038</td>
<td>0.155</td>
<td>0.061</td>
<td>0.016</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Failed CAHSEE ELA in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>CST ELA</th>
<th>CST Math</th>
<th>GPA</th>
<th>% Absent</th>
<th>Learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed Math</td>
<td>-0.012</td>
<td>-0.170</td>
<td>-0.053</td>
<td>0.419</td>
<td>-0.022</td>
</tr>
<tr>
<td></td>
<td>(0.105)</td>
<td>(0.114)</td>
<td>(0.130)</td>
<td>(1.632)</td>
<td>(0.096)</td>
</tr>
<tr>
<td>Observations</td>
<td>543</td>
<td>497</td>
<td>539</td>
<td>543</td>
<td>543</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.071</td>
<td>0.036</td>
<td>0.022</td>
<td>0.010</td>
<td>0.060</td>
</tr>
</tbody>
</table>

Passed CAHSEE ELA in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>CST ELA</th>
<th>CST Math</th>
<th>GPA</th>
<th>% Absent</th>
<th>Learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed Math</td>
<td>0.099</td>
<td>0.093</td>
<td>-0.040</td>
<td>1.854*</td>
<td>-0.124**</td>
</tr>
<tr>
<td></td>
<td>(0.076)</td>
<td>(0.081)</td>
<td>(0.115)</td>
<td>(0.710)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,222</td>
<td>1,153</td>
<td>1,210</td>
<td>1,222</td>
<td>1,222</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.305</td>
<td>0.013</td>
<td>0.023</td>
<td>0.005</td>
<td>0.084</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p < 0.01, * p < 0.05; Bandwidth=4
Table A.2.13: Robustness Checks: Testing for Breaks in the Control Variables in Grade 10 - Ethnicity and Gender

Failed Math in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Female</th>
<th>American</th>
<th>Hispanic</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed ELA</td>
<td>-0.070</td>
<td>-0.089</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>(0.084)</td>
<td>(0.080)</td>
<td>(0.072)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>Observations</td>
<td>798</td>
<td>798</td>
<td>798</td>
<td>798</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.002</td>
<td>0.039</td>
<td>0.015</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Passed Math in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Female</th>
<th>American</th>
<th>Hispanic</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed ELA</td>
<td>-0.051</td>
<td>-0.033</td>
<td>0.075</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>(0.080)</td>
<td>(0.069)</td>
<td>(0.082)</td>
<td>(0.059)</td>
</tr>
<tr>
<td>Observations</td>
<td>698</td>
<td>698</td>
<td>698</td>
<td>698</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.010</td>
<td>0.013</td>
<td>0.008</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Failed ELA in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Female</th>
<th>American</th>
<th>Hispanic</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed Math</td>
<td>0.031</td>
<td>0.011</td>
<td>-0.032</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>(0.088)</td>
<td>(0.079)</td>
<td>(0.139)</td>
<td>(0.088)</td>
</tr>
<tr>
<td>Observations</td>
<td>543</td>
<td>543</td>
<td>543</td>
<td>543</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.016</td>
<td>0.007</td>
<td>0.006</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Passed ELA in Grade 10

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Female</th>
<th>American</th>
<th>Hispanic</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed Math</td>
<td>0.216**</td>
<td>0.123</td>
<td>-0.033</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.061)</td>
<td>(0.079)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,222</td>
<td>1,222</td>
<td>1,222</td>
<td>1,222</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.030</td>
<td>0.007</td>
<td>0.018</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, clustered at the CAHSEE score combination level

** p< 0.01, * p< 0.05; Bandwidth=4
2.11 Acknowledgements

I thank my co-author and chair of my doctoral committee, Professor Julian R. Betts for all the valuable comments and suggestions. I would like to extend further thanks to Karen Volz Bachofer and Andrew Zau for the technical and institutional support and to the San Diego Unified School District for the data used in the analysis.

Chapter 2 is currently being prepared for submission for publication of the material. The dissertation author was the primary author of this material. Zieleniak, Jedrzej; Betts, Julian R. “Effects of Test Failure on Academic Achievement in the Context of Student Accountability: A Regression Discontinuity Analysis.”
Chapter 3

Chasing the Syllabus: Measuring Learning Trajectories in Developing Countries with Longitudinal Data and Item Response Theory

Karthik Muralidharan$^1$ and Jedrzej [Yendrick] Zieleniak$^2$

---

$^1$University of California, San Diego, Department of Economics, 9500 Gilman Drive, La Jolla, CA 92037, USA. E-mail: kamurali@ucsd.edu. Web page: http://econweb.ucsd.edu/ kumaridharan/

$^2$University of California, San Diego, Department of Economics, 9500 Gilman Drive, La Jolla, CA 92093, USA. E-mail: jzieleni@ucsd.edu. Web page: http://econweb.ucsd.edu/ jzieleni/
3.1 Abstract

Combining a unique annual longitudinal dataset on learning outcomes in rural public primary schools in the Indian state of Andhra Pradesh (AP) with a vertically-scaled item-response theory (IRT) model, we present the first estimates of learning trajectories for a cohort of students in a developing country over the course of five years of primary school. We make four main contributions in this paper: (1) We provide the first out of sample validation of a vertically-scaled IRT model in a low-income setting and show that the model can reliably be used to measure learning trajectories over time; (2) In addition to confirming that learning levels are very low, we also find that the average rate of learning in each school year is significantly less than the rate of progress required to keep pace with the syllabus in other words the average rural public school student falls further behind grade-appropriate standards each year (especially for math); (3) We find a significant increase in inequality of learning outcomes as students go through primary school, with increasing disparities by socio-economic status over time; (4) We find that the absolute productivity of an additional year in school beyond second grade is very low with less than 15% of students who cannot answer a question at the end of second grade being able to do so at the end of an extra year of schooling, suggesting that students who fall behind in early years learn very little from additional years in school. Our results suggest that the syllabi and pedagogy of education systems in developing countries have not effectively made the transition from serving a screening function to being able to add human capital at all parts of the student achievement distribution. They may also partly explain why increases in business as usual school inputs in the past decade may not have led to much improvement in learning outcomes in developing countries.

Keywords: Education Policy, Learning, Human Capital, Item Response Theory
3.2 Introduction

Progress in many areas of human well-being relies on having well defined and reliable cardinal measures of outcomes, that policy can then seek to improve. Examples of such measures include per capita income, poverty rates, infant mortality rates, and measures of stunting and wasting. Cardinal measures of education to date have included literacy rates, school enrollment and completion rates, and the average years of schooling completed (Barro (1991); Mankiw, Romer, and Weil (1992)), and international development goals for education have typically focused on these measures\(^3\). Such an approach implicitly assumes that a year of education in one country is equivalent to one in a different place.

However, an increasing body of evidence suggests that this assumption may be highly misleading, with substantial variation across countries in the extent to which an additional year of schooling produces cognitive skills and human capital (Schoellman (2012); Singh 2014). Further, recent evidence suggests at both the macro and micro levels that what matters for both growth as well as employability are not years of education as much as the quality of education represented by learning outcomes and skills (Nickell (2004); Hanushek and Woessman (2007); Schoellman (2012)). Finally, while many developing countries have made substantial progress in improving school enrollment rates, independent assessments of learning show that learning levels are extremely low in countries ranging from India (ASER (2013)) to Kenya, Uganda, and Tanzania (UWEZO (2012)).

Taken together, these results highlight the importance of documenting and better understanding how educational systems perform in converting a year of schooling into human capital (Das and Zajonc (2010)). Nevertheless, even when they exist

---

\(^3\)The Millenium Development Goals (MDG’s) for education is for all children to complete primary education. The Human Development Index (HDI) measures education using adult literacy and school enrollment rates.
(which is not often), measures of learning levels in developing countries are based on snapshots in time, and do not track the dynamics of learning for a cohort over time, which is a key limitation in understanding the performance of education systems. The absence of data on learning trajectories reflects both a lack of longitudinal data on test scores for the same students over time, as well as methodological challenges in constructing and equating tests to establish cardinal measures of learning levels that can be used to measure learning gains over time.

We address both these difficulties in this paper by combining a unique annual longitudinal dataset on learning outcomes collected over five years in a representative sample of 100 rural primary schools in the Indian state of Andhra Pradesh (AP), with a vertically-scaled item-response theory (IRT) model that enables out of sample prediction of student performance on a representative sample of grade appropriate questions. First, we use data from a specially designed cross-sectional assessment conducted among over 6,000 students in these 100 schools that allows us to test the out-of-sample validity of a vertically scaled item-response theory (IRT) model. Second, we use annual data on test scores collected for over 20,000 unique students in the same 100 schools over a 5-year period to estimate IRT item parameters for a question bank of over 1,500 test items in math and language.

When combined with annual longitudinal data on test scores, the model and the item parameters allow us to estimate the evolution of a cardinal measure of student ‘ability’ over time, and to characterize the evolution of student ability at different percentiles of the achievement distribution\(^4\). Further, combining the estimates of student ability over time with the IRT item parameters of questions corresponding to

---

\(^4\)Note that the term ‘ability’ in IRT models does not refer to an intrinsic non-changing measure of ability as in many economics models. Rather, it refers to a measure that is correlated with test scores, and can be thought of as a proxy for a cardinal measure of human capital. Importantly, the IRT models allow us to estimate the ‘ability’ score for the same student at different points in time and thereby trace the evolution of a proxy for human capital as students go through the school system. Caveats are discussed in more detail in Section 2.
different grade levels of difficulty allows us to characterize the extent to which student learning levels (proxied by their ability score) are evolving relative to the standards envisaged by the syllabus. Thus, the central contribution of this paper is to validate a standard psychometric model (vertically-scaled IRT) in a low-income setting and combine the model with longitudinal data to estimate learning trajectories (and their distributions) for a representative sample of rural public school students over the course of five years in primary school. We present four main results in this paper.

First, we present the first out of sample validation of the use of a vertically-scaled IRT model for measuring learning progress over time in a low-income setting and show that the model performs very well in such an exercise. The correlation between the predicted and actual scores at the classroom level is 0.95 in math and 0.8 in language, and the predicted score is able to explain 85% of the variation in actual scores in mathematics and 65% of the score variation in language. We also show that the model performs well within sample in the longitudinal data, where predicted scores explain around 90% of the variation in actual student scores in math and in language. These results provide confidence in using the model and the estimated item parameters to trace learning trajectories over time.

Second, we corroborate evidence from several other studies (including Pratham ASER (2013) and UWEZO (2012)) that learning levels in low-income settings are very low in spite of high primary school enrollment rates. Less than 10% of fourth and fifth grade students would score 50% or more on a grade appropriate math test and less than 40% would do so in language. Strikingly, only 25% and 55% of students in grade 5 would score over 50% on a grade 2 test in math and language respectively. Moving beyond levels to trajectories, we find that the median student’s learning trajectory is significantly flatter than the rate of growth envisaged by the syllabus, implying that the typical student falls further behind the expected grade-appropriate standard for
each additional year in school. This is especially true for math.

Third, we find a significant increase in the variance of student ability scores over time, and the gap in test scores by socio-economic status (SES) of students grows over time. These results suggest that inequalities in learning levels among students may be growing even as inequalities in primary school enrollment and completion rates are falling. Strikingly, we estimate that the lowest decile of students may be learning nothing in math after grade two even after spending an extra three years in school. Thus official data on reduced school drop out rates and increased retention may be inducing a misleading sense of reducing inequality in education attainment by masking the increase in inequality in learning outcomes.

Fourth, we show that the absolute productivity of an extra year in school after second grade is very low. Specifically, we show that for any given number of students who are not able to answer a first or second grade question at the end of second grade, less than 10% of them will be able to do so after an extra year in school. Graphical plots of abilities over time reveal a sharp flattening of learning trajectories beyond second grade, which suggests that the value of an additional year in school beyond second grade may be very low if students are not able to read adequately by then.

Our results fill an important gap in the literature on education in developing countries, and are likely to have direct policy implications. While a growing body of high-quality experimental evidence in the past decade has helped identify the impact of specific interventions on education outcomes in low-income countries\textsuperscript{5}, there has been almost no credible documentation of the ‘business as usual’ performance of schooling systems in low-income settings as a whole. As Figure 3.1 (Panel A) shows, the literature has focused on credibly estimating the difference between treatment

\textsuperscript{5}Examples include Banerjee et al. (2012), Muralidharan and Sundararaman (2011) and Muralidharan and Prakash (2013), and Duflo et al. (2012); see Kremer et al. (2013) for a comprehensive review.
and control groups without building a corresponding understanding of how the control group itself is evolving over time. This is a striking gap because schooling is the most expensive investment that societies and individuals make in the accumulation of human capital. Further, understanding the extent of improvement in a ‘business as usual’ setting allows the test score impacts of any intervention to be measured in terms of the equivalent ‘years of schooling’.

FIGURE 3.1 HERE

While the trajectories we estimate obviously cannot be interpreted as the ‘causal’ impact of schooling on learning outcomes, they are likely to constitute an upper bound on the impact of schooling on learning outcomes given that the time period of our data (2005-2010) is a period of economic growth and improving incomes that are likely to be positively correlated with learning outcomes. More broadly, trajectories of the sort we estimate allow for monitoring of both the quality (mean achievement) as well as the equity (variance of achievement) of an educational system over time. The analysis of the distribution of learning outcomes may be a particularly important barometer of social progress in reducing inequalities in human capital.

The rest of the paper is organized as follows. Section 2 discusses the use of vertically-scaled models to estimate learning trajectories, presents our approach to model validation and discusses the data we use; Section 3 presents results from validating the vertically-scaled IRT model in our setting; Section 4 presents our main results on learning levels, gaps, inequality, and productivity of the school system. Section 5 discusses the results, caveats, and policy implications, and concludes.
3.3 Estimating Learning Trajectories

3.3.1 Methodological Issues

Figure 3.1 (Panel B) provides a simple sketch of the objective of our exercise. The 45-degree line represents the rate of progress in human capital accumulation that is envisaged by the syllabus and embodied in the textbook. In practice, it is well known that many students in low-income settings exhibit much lower learning levels than what would be a grade-appropriate standard (Pratham 2013; Uwezo 2012). Thus, it is likely that the rate at which student achievement is growing is considerably below this line (as illustrated in the Figure). However, while the conceptual idea behind Figure 1 (Panel B) is straightforward (see Pritchett (2004)), we are not aware of any work to date that has been able to populate this conceptual figure with real numbers. The lack of evidence on learning trajectories in low-income settings reflects several challenges.

First, there are very few data sets that document cardinal progress in learning levels over time. While measurement exercises such as those conducted by the ASER reports in India, the Uwezo reports in East Africa, and the PISA scores do report average population learning levels at different points in time, these are based on repeated cross-section snapshots and do not follow the same students over time.

Second, even having longitudinal data on student test scores is not enough by itself, because we also need a conceptual framework to define a cardinal unit of measurement to document progress in learning outcomes over time. A simple approach would be to ask the same question to students over time in school and code the vertical axis in Figure 3.1 (Panel B) as the probability of a typical student getting the question right. In practice, such an approach has several limitations. First, constraints on test length limit how many overlapping questions from previous years can
be included (especially spanning several grades). Second, and more importantly, such an approach is (a) subject to bias - due to acclimatization of students to the repeated questions over time, and (b) can also be easily gamed by teaching the repeated questions to the exclusion of other questions, which can give a highly misleading estimate of the amount of progress that students are making over time. Third, even if the numbers on average fraction correct on a question over time were to be reliable indicators of human capital accumulation (which is unlikely), it is difficult to interpret these numbers beyond the specific questions asked and to infer much about distributions of learning.

Thus, an essential focus of the field of psychometrics has been to address questions of test design, reliability, and comparability that enable the scoring and comparison of student abilities in ways that are robust to gaming and enable comparisons across students who do not take identical tests. One of the most widely used techniques in standardized testing is that of item-response theory (IRT), which is what we use in this paper.

### 3.3.2 Item-Response Theory (IRT) models

The core idea of IRT models is that there is an underlying student 'ability' at a point in time that can be represented on a unidimensional scale, and that test questions (items) have item-specific characteristics (identified by the ICC or the Item Characteristic Curve) such as their 'difficulty' and their 'discrimination' (see Appendix 1 and also Das and Zajonc (2010) for non-technical summary of IRT). IRT models in practice aim to construct a large question (item) bank where the IRT-parameters of each item (the ICC) are calculated in a reference population. Once such a question bank is created, it is possible to create different tests that have comparable properties by suitably sampling questions that may be different but have
similar ICC’s.

The main strength of the IRT is the fact that the estimated models are not sample or test dependent, which provides flexibility when different tests are administered to different samples of students. The test and sample independence implies that the model can be used to estimate the probability that a student with a given ability score can answer a question correctly, *even if the student never attempts that particular question*. While many aspects of the IRT framework remain debated in the psychometric literature, the use of the approach itself is widely accepted, and used in almost all large-scale high-stakes assessments of student academic aptitude - including the Scholastic Aptitude Test (SAT), the Graduate Record Examination (GRE), and the Graduate Management Admission Test (GMAT).

IRT models enable two types of comparisons of students. The first is the comparison of ‘ability’ across students against a fixed domain of competence. This is achieved by a technique referred to as ”horizontal scaling” and is used in all the cases mentioned above (SAT, GRE, and the GMAT). While the ‘ability’s score and scale are completely arbitrary, it is easier to interpret the scale because the main aim of the exercise is to rank students against each other on their master of a fixed set of competences. Thus in practice, scores are reported on a standardized scale (of say 200-800) where the end points of the range represent the highest and lowest scoring students in the reference distribution.

3.3.3 Vertically-Scaled IRT models

A second kind of comparison is the evaluation of student progress across grades, which is achieved by a technique referred to as vertical scaling. The scoring algorithm works in exactly the same way as in horizontal scaling, with the main difference being that students of different ages and grades are pooled together in the scoring, and
the same student at different points in time constitutes distinct observations. The changes in ‘ability’ score of the same student over time can then be treated as a cardinal measure of human capital. Figure 3.1 (Panel A) provides a sketch of the approach, with the key quantity of interest being the extent to which the ability distribution of a cohort of students shifts to the right over time.

Note that this scale also has no intrinsic meaning by itself (as with horizontal scaling). However, combining the estimates of the changes in ability score of a given student over time with the ICC’s of specific questions allows us to estimate the change in the probability of that student getting any given question correct over the same time period (even if that student did not answer that question in any of the periods). Averaging across all the students in a cohort, the approach can then be used to estimate the change in probability of students in the cohort getting a particular question correct. Similarly, averaging the performance of a cohort over time on all grade-appropriate questions in the question bank would allow us to estimate the change in probability over time of a cohort being able to answer questions at different grade-levels, which in turn would allow us to estimate learning trajectories for an entire cohort.

However, vertical-scaling is significantly more contentious than horizontal scaling, as the domain of knowledge covered in each test changes as the student moves through the educational system (unlike in the SAT or the GRE). Furthermore, the decisions about the choice of a model and implementation procedures might affect growth interpretations (Briggs et al. (2008a)). The shifting construct domain requires sufficient items to be common across tests in order to allow for reliable model estimation. The psychometric literature suggests that a 20% question overlap between adjacent grades may be required to credibly estimate vertically-scaled IRT models (Kolen and Brennan (2004)). However, such a large overlap across tests over time
could in turn compromise the integrity of the scaling exercise (due to familiarity and gaming as discussed earlier).

Thus, a key contribution of our paper is to conduct an independent out-of-sample validation of a vertically-scaled IRT model using a cross-sectional validation exercise. To our knowledge, the out of sample validity of vertically scaled IRT models has not been a widely investigated topic, as the literature has focused on model selection rules instead. Our review of the psychometric literature suggests that ours is the first study to validate a vertically scaled model out-of-sample in a low-income country context.

### 3.4 Validating the vertically-scaled IRT model

#### 3.4.1 Context and Data

India has the largest primary schooling system in the world, catering to over 200 million children. This paper uses data from the Indian state of Andhra Pradesh (AP), which is the 5th most populous state in India, with a population of over 80 million (70% rural). The data was collected as part of the Andhra Pradesh Randomized Evaluation Studies (AP RESt), a series of experimental studies designed to evaluate the impact of various input and incentive-based interventions on improving education outcomes in AP\(^6\). The project collected detailed panel data over five years (covering the school years 2005-06 to 2009-10) on students, teachers, and households in a representative sample of 500 government-run primary schools (grades 1 through 5) across 5 districts in Andhra Pradesh. The dataset includes annual student learning outcomes as measured by independently conducted and graded tests in language (Telugu) and math (conducted initially at the start of the 2005-06 school year as a

---

\(^6\)These interventions are described in Muralidharan and Sundararaman (2011)
baseline, and subsequently at the end of each school year), basic data on student and teacher demographics, and household socio-economic data for a subset of households.

Note that all estimates of learning trajectories in this paper use only the data from the 100 control schools that represent the business as usual progress of students through the school system, and do not use data from any of the treatment schools. The data encompasses 39,103 student-year observations in math and 39,162 in Telugu, with 20,005 unique student IDs present in the data set. The distribution of data across grades and years is presented in Table 1. The oldest cohort of students is the one in grade 5 in Year 1 (2005-06), while the younger is the one in grade 1 in Year 5 (2009-10). The size of each cohort entering the public schools during the study period decreases over time, which most likely reflects the increasing extent to which students in younger cohorts enroll in private schools.

However, we also see that the number of observations over time for a given cohort are pretty constant (as seen in the diagonal elements of Table 3.1 for any cohort), suggesting that differential selection into private schools may mostly be taking place prior to grade 1\(^7\). Since most of our focus is on learning trajectories conditional on starting in public school, and almost all our analysis includes cohort fixed effects, our main results are unlikely to be affected by differential selection into private schools over time. Nevertheless, it is important to note that the tests were conducted in school and not in the household, and so our data represent the learning trajectories for a representative sample of students who attend public schools, and not a sample of all students.

\(^7\)This is consistent with evidence in Muralidharan and Sundararaman (2013) where they show that most private schools also have 2 years of pre-kindergarten and kindergarten, whereas public schools start in grade 1. Thus, the decision to enroll in private school is typically taken before grade 1.
The assessments were created, extensively piloted, and administered and scored by an independent agency, ensuring that the test scores are valid measures of learning and their scores are not biased by teacher subjectivity. In addition to being used to estimate the impacts of various treatments, the test design also included overlapping questions over time and across grades to enable the construction of a vertically-equated scale to estimate learning trajectories over time. However, since the tests were high-stakes for some of the treatment groups (such as the one where teachers received bonus payments based on the improvement of their students test scores), the number of common (linking) questions was limited, since the introduction of significant amount of item overlap between tests would provide the teachers with the option to game the testing regime, significantly biasing the test scores. Further, even if the number of overlapping questions was high, it would not have been possible to conduct an out of sample validation of the vertically-scaled IRT model (as specified below).

3.4.2 Linking, Equating, and Validation Design

We therefore conducted an additional Special Assessment, after the five-year study period in 2012, which served two purposes: first, it helped address the insufficient number of linking questions across grade levels in the main data set, and second, it enables us to validate the vertically scaled IRT model out-of-sample. The assessment included all students attending the 100 public primary schools in rural Andhra Pradesh which previously served as a control group for the experiments carried out within the AP RESt research program, yielding 6,152 student level observations, distributed across grades one through five.

There are several important features in the construction of this Special As-
essment. First, this test was to serve only diagnostic purposes so the potential for teachers to respond strategically to the test was low, and allowed us to design tests with significant item overlap across grade levels. Second, the linking items (common questions) were administered cross-sectionally across students in different grades at the same time, and thus the estimates of item parameters will not suffer from the familiarization bias that may take place if students take the same question multiple times. Third, and most importantly, the test construction allows for validation of the vertically-scaled IRT model by testing its predictions out of sample (as shown below).

The test design and the details of item overlap across grades are presented in Table 3.2. Students in grades one to five were tested on 25, 50, 60, 70, and 70 questions each on math and language respectively. The questions were divided into common questions (categories A-D in Table 2) and adjacent questions (E-I). The adjacent questions were introduced at the grade appropriate level (the 15 questions in group E correspond to the grade one level, the 15 in group F to grade two level, and so on) and included in the exam administered to the appropriate grade and the grade above. A standard vertically-linked testing protocol would use sets of adjacent questions to score all the items and put them on a common scale, and the adjacent questions mimic such a protocol.

TABLE 3.2 HERE

The main innovation in our approach is the use of common questions (A-D) that were introduced at the grade appropriate level and included in the exams administered to all the students at that grade level and at all higher grades. Thus questions in group A are at the same difficulty level as those in group E, and are so
those in groups B and F, C and G, and D and H (each comprising questions at the level of grade one, two, three, and four respectively). However, the common questions let us observe the actual score that a typical student in grades three, four or five will obtain on questions of all previous grades.

Since standard vertically-scaled IRT models only use overlapping questions between adjacent grades, and do not have common questions of the sort we have, we treat the common questions as adjacent in the estimation of vertical scaling of the IRT model. When estimating the model, we omit student responses to grade 1 questions in grades 3-5 (A3-5), responses to grade 2 questions in grades 4-5 (B4-5) and responses to grade 3 questions in grade 5 (C5). The data omitted form the estimation is highlighted in gray in Table 3.2, and this data is used to validate the model estimated using the rest of the data. Specifically, we estimate the model using the non shaded data in Table 3.2, and then use the model, combined with estimated ability parameters for students in grades three to five and the ICCs of the questions in groups A-C to predict the probability of a correct response to the shaded questions. We then compare these predictions to the actual responses (which were omitted from the estimation).

The literature on implementation of IRT models in education suggests the use of questions from both lower as well as upper grades as adjacencies (Reckase (2010)). However, in our context, patterns of student responses at grade appropriate levels are often very low. Introducing a question appropriate at grade N-level to grades lower than N would not be very useful since the majority of students would score zero correct responses. Thus, introducing links from the upper to lower grade would not provide any useful information for the IRT estimation. Instead, we take advantage of the variation in the question difficulty of questions administered at any given grade level.
In addition to the ability to validate a vertically-scaled IRT model out of the estimation sample (using the common questions), the Special Assessment also allows for a more robust construction of a vertically-linked item scale because we now have 25 overlapping questions between each adjacent grade. Further, the 115 questions (in each of math and language) were a subset of the universe of nearly 1500 questions that were used in the longitudinal data over five years in the main study, and these 115 questions in turn were sampled to be representative of the master question bank (stratified so that questions from all five years and five grades were represented).

The special assessment data was used to estimate a vertically scaled IRT model with data exclusions as specified in Table 3.2. For all common questions introduced in grades one, two and three we exclude student answers in grades more than one grade level above the grade in which the question was first asked. We estimate a concurrent 3 parameter logistic (3PL) model using all five grades. We use the IRT estimates to predict the probability of a correct answer for each student and each question asked in the Special Assessment. The exclusion of responses for 30 questions in grades three through five in the IRT estimation implies that the estimated probabilities of correct answers were obtained without taking into account the actual student responses to those questions. We limit the analysis of the validity of the vertically scaled 3PL IRT model to the observations not used in the model estimation, encompassing 3940 student observations. Special assessment test design implies that the out of sample predictions are available for a set of 10 questions for 1,303 students in grade three, a set of 20 questions for 1,270 students in grade four, and a set of 30 questions for 1,367 students in grade five.
3.4.3 Results: Validation of vertically-scaled IRT in cross-sectional data

Since the answers to individual questions are binary, we validate the vertically-scaled IRT model by comparing the predicted and actual mean scores for each student in grades three to five (in the Special Assessment sample) on the set of common questions that were not used for the estimation (these are the shaded questions in Table 3.2). As a validation exercise, we first report the correlations between the predicted and actual scores, and then examine the distribution of the prediction errors.

Table 3.3 presents the correlation between student i’s actual ($UQN_i^{SCR}$) and predicted scores ($IRT_{Pred,Scr_i}$) using Equation 3.1 with robust standard errors clustered at the school level. The results are presented at the student level, the school-grade level, and the district-grade level because these are the policy-relevant levels at which we may want to know how performance is evolving over time.

$$UQN_i^{SCR} = \beta_0 + \beta_1 IRT_{Pred,Scr_i} + \varepsilon_i$$  (3.1)

TABLE 3.3 HERE

The regression results suggest that the IRT model estimates perform extremely well in math with a coefficient over 0.96 on predicted math score for all observation levels, with all being statistically significant at the 1% level, suggesting a high correlation between the observed score and those predicted by the model (the raw correlation at the student level is 0.92). The math predicted score explain between 85 and 90 percent of the variation in the actual scores at the student, school-grade, and district-grade levels. The Telugu (language) predictions are less precise, but the model still does quite well and predicts around 65, 75, and 85 percent of the total variation at
the student, school-grade, and district-grade levels, with the regression coefficients in
Equation 3.1 estimated for Telugu ranging from 0.72 at the student-level to 0.88 at
the district-grade level (all coefficients are again significant at the 1% level).

Graphical comparison of the distribution of actual student performance and
the model predictions for math and language confirms previous observations of a
better model fit in math (Figure 3.3). Figure 3.3 presents the comparison of the
distribution of the predicted scores based on the IRT model and the actual pattern
of responses for the out of sample question pool by subject. Formal tests of goodness
of fit with a large number of observations are predicted to fail (Browne and Cudeck
(1993)), and the literature does not provide a viable alternative to the conventional
tests of equality of distributions. In order to compare the distributions of actual and
predicted scores we carry out the chi-squared goodness of fit tests, the Kolmogorov-
Smirnov tests for equality of the distributions and the two-sample mean comparison
t-tests in each of the out of sample grades separately, and find that in most cases, we
do reject the null hypothesis of equality of the distributions.

However, while the formal tests reject the equality of the distributions between
the observed and the predicted scores, the regression results in Table 3.3 and the dis-
tributions of errors of the predictions suggest that the vertically scaled IRT model
performs well in predicting student scores. Overall, the results of the validation ex-
cercise suggest that the model is likely to provide a highly accurate representation of
learning trajectories in mathematics, and that while the predictions in language may
be less accurate at the level of the individual student, they appear to be accurate
enough when aggregated up to the school-grade or district-grade level to provide use-
ful policy-relevant measures of tracking learning trajectory at the school and district

---

8The psychometric literature suggests an Adjusted χ²/df test, but simulation studies show that
with small number of questions or a small number of observations the test performs poorly (Tay and
Drasgow (2012)).
3.4.4 Results: Validation of vertically-scaled IRT in longitudinal data

The Special Assessment data encompassed 230 questions (115 questions each in math and language). As discussed earlier, all these questions were taken from the master database of over 1,500 questions in math and language, with stratified sampling across years and grades. The item parameters estimated in the special assessment were used as fixed parameters in the estimation of a vertically scaled IRT model in the five year data. The model was estimated using 39,103 student-year observations in math and 39,162 student-year observations in language. The model provides estimates of ability parameters for each student at each point in time, and combined with the estimates of the item parameters of the full set of over 1500 questions, we are able to predict the probability of each student in the sample getting any question in the database correct at any point in time. Averaging the predicted scores on each each actual test provides a predicted score for each student in the longitudinal data on the test that she actually took. As in the previous section, we can assess the validity of this model-based prediction exercise by comparing the predicted score with the actual score.

We investigate the predictive power of the full model using the regression model specified in Equation 3.1. The model is estimated at the student-year, grade-school-year and grade-district-year level and the regression results are reported in Table 3.4. The within-sample predictive power of the full model is very high. Predicted math scores explain 89.7 percent of the variation in actual scores at the student-year level, 96.4 percent of the variation at the grade-school-year level and 98.2 percent of the variation at the grade-district-year level. Predicted language scores explain 91.8
percent of the variation in the actual scores at the student-year level, 97.1 percent of the variation at the grade-school-year level, and 99.5 percent of the variation at the grade-district-year level. Furthermore, the coefficients on the predicted scores are statistically significant at the 1% level in every specification. The model performs very well within the longitudinal data used in the analysis.

Another way of assessing the performance of the model is to plot the prediction errors (defined as the difference between the actual and predicted scores at each level) at the student, grade-school and grade-district level. We do so in Figure 3.4 and see that while prediction errors at the student-level are large enough to be cautious about making student-level predictions, the prediction errors are considerably lower at the school-grade level (a 95% confidence interval ranging from -0.018 to 0.033 for math and -0.02 to 0.025 for language), and even lower at the district-grade level. Since our objective is not to make student-level predictions but to assess the performance of the education system over time, where we will pool all the data from a given cohort and track the evolution of performance over time, the tight distribution of the prediction error at the school-grade and district-grade levels suggest that the vertically-scaled IRT model performs well enough to be used for this purpose.

FIGURE 3.4 HERE

3.5 Learning Trajectories

3.5.1 Levels

Having validated the vertically-scaled IRT model, we now present the main results on learning trajectories that such an approach allows us. First, the model allows us to estimate an ability score for each student at each point in time. Plot-
ting the distribution of ability scores at different grades and observing the extent to which the distribution shifts to the right over time will be equivalent to populating the conceptual Figure 3.1 (Panel A) with real data. In particular, we focus on the cohort that we follow for its entire time in primary school over five years (we refer to this cohort as cohort 5).

Figure 3.5 plots the kernel densities of the ability score parameters in math and language for the 5-year cohort observed in the longitudinal data. In both panels the distributions associated with ability parameters of students in grade 5 are shifted considerably to the right relative to the distribution of the students form the same cohort in grade 1, four years earlier. Thus, the data suggest that the students are in fact learning while in school. The data further shows that large masses of students in both subjects are concentrated in the lower range of the ability scores. This observation reflects the fact that many students enter primary school with very low levels of academic abilities. However, as they progress through the school, the data shows that the large masses in the left tails shift towards the middle of the distribution, further suggesting that students are, in fact, gaining from attending school.

FIGURE 3.5 HERE

However, the underlying expectations of the syllabus are also changing in this period. The ability to divide data according to the grade appropriateness of the questions asked allows us quantify this change in expectations over time. For each grade we use the universe of grade appropriate questions in order to estimate the ability parameter which would allow a student to obtain a score of 50% in the entire universe of grade level items. We consider that minimum ability level as the grade appropriate standard. The choice of the cutoff is arbitrary; the same exercise could be carried out
using other cutoff scores. We use the construct of a grade-appropriate standard to investigate how students learn, relative to where they would potentially be expected to be at a given grade level, which in Figure 3.5 is captured by the vertical dotted lines. The locations of these vertical lines (marking standards) already shows that most students are far behind where they are expected to be, but the point will be made more explicit in later parts of the analysis.

Since the ability scale shown in Figure 3.5 is arbitrary, the most concrete way of interpreting these abilities is to use the ability scores and item-parameters for grade appropriate questions to calculate the mean probability students in a given grade getting a question correct (of that grade or an earlier grade). We shift our attention to the probabilities of a correct answer on a grade-appropriate question. The predicted probabilities allow us to calculate what fraction of students would score at above 50% on any grade level test, and at each grade level. Table 3.5 presents the fraction of students predicted to score above 50% in each grade, on a test at each grade-level.

TABLE 3.5 HERE

The data shows that the fraction of students doing well in mathematics on the grade-appropriate material, as we move down the main diagonal (i.e. scoring 50% on more on a grade 1 test in grade 1, grade 2 test in grade 2, etc.) is very low. Only 2.44% of students in the 5-year cohort is predicted to score at or above the benchmark in grade 1, while 16.05% manages to meet the standard in grade 2. In language student learning levels appear to be much better– 11.45% of first graders meet the standard in grade 1, while 28.02% meet the grade 2 standard in grade 2. The higher fraction of students meeting the language standards as seen in Table 3.5 can also be observed in Figure 3.5, as a larger mass of students are to the right of the grade 1
and grade 5 standards, as compared to math.

We can further investigate how students would perform if we kept the testing content the same at all grade levels. Each row in Table 3.5 corresponds to a given grades difficulty level. Each column focuses on students at a given grade level. Table 3.5 shows that in the case of all grade appropriate tests, the students are predicted to do better, as they move to a higher grade. While only 2.44% are predicted to meet the grade 1 standard in grade 1, by grade 2 36.24% of students at that grade level are predicted to meet the grade 1 standard. In both math and language, the largest gains are associated with the grade 1 material, especially as the students move from grade 1 to grade 2. In the case of material associated with grades 2 through 4 the increases in fraction of students reaching the standard as the students move through grades is significantly lower, suggesting flatter learning trajectories than in the case of grade 1 material.

We further investigate the learning levels and learning trajectories specific to each grades material graphically. Figures 3.6 through 3.9 present the learning trajectories for grade 1- through 4- appropriate material, encompassing observations for all cohorts present starting with the grade where the material is first introduced, all the way until grade 5. Each figure presents the trajectories for the 10th, 25th, 50th, 75th and 90th percentile of the grade-year ability score distribution, along with the dotted horizontal line at 50% representing the given grade-levels standard.

FIGURE 3.6 HERE

Figure 3.6 reveals that the median student reaches the 50% score benchmark in language by grade 3, while in math it take the median student until grade 4 to reach the grade 1 standard. The trajectories suggest that the level of school readiness
is lower in math than in language, which can be interpreted as the reflection of the linguistic learning happening through daily interactions before entering school and the lack of equivalent informal education in the case of math.

Figure 3.6 further shows that we can observe a steeper trajectory of learning between grades 1 and 2 than between other grade levels. For the students at the 10th percentile of the grade-level ability distribution, the trajectories of learning after grade 2 are flat, suggesting that those students gain nothing from attending school, after the initial gain between grades 1 and 2. In other parts of the distribution, the same phenomenon can be observed. The trajectories are steeper between grades 1 and 2 and become flatter between grades 2 through 5. The likely reason for the steeper trajectory is the fact that grade 1 questions are being read to students, while beyond grade 2 the students have to be able to read in order to answer test questions.

The grade 1 material learning trajectories suggest that on average students learn as they progress through primary school. However, when we shift the testing domain to grade 2, 3 or 4 material, the learning levels become almost completely flat, for all parts of the achievement distribution.

FIGURE 3.7 HERE
FIGURE 3.8 HERE
FIGURE 3.9 HERE

We can still observe very small positive gradient of the trajectories associated with grade 2 material, as presented in Figure 3.7, but beyond grade two material, the additional schooling does not contribute to student learning.
3.5.2 Inequality

The quality of the educational system, as measured through learning levels, is one of the main concerns of policy makers. Another major concern is the system’s equity and whether additional years of schooling narrow or widen the achievement gap between high and low-performers. We investigate whether the achievement gaps between students increase as they progress through primary school by focusing on grade specific variances of the ability score parameters for the 5-year cohort. We report results of the pairwise Leven’s tests for equality of variances in Table 3.6.

TABLE 3.6 HERE

The first row in each panel, math and language reports the grade specific variances of the ability score parameters. The main diagonal presents the Leven’s test statistics for adjacent grades, while the remaining entries correspond to the test for the variances of the row’s and column’s grades. The test results show that for the tests of adjacent grades’ variances, the null hypothesis of the equality of variances can be rejected in six out of eight cases. It is important to notice that in both math and language the variances of ability score parameters go down between grade 1 and 2 and rise from then onwards. The fall between grades 1 and 2 suggest that the initial year of schooling helps to catch some of the students up, at least to some extent. However, as the remaining tests show, as the students move through grade levels, the variances increase in both subjects, suggesting that educational inequalities increase with schooling.

Educational equity concerns in the context of India further extend to the inequalities stemming from socioeconomic status. The analysis presented so far focused on student learning and inequalities without consideration for student background and
household characteristics. However, in a context where social stratification based on caste might be of potential concern, the investigation of heterogeneity in learning based on socioeconomic status might shed further light on educational inequalities. We use the household surveys in order to construct student’s socioeconomic (SE) index. We investigate the potential socioeconomic-based inequalities in learning by investigating the ability score parameter ($\theta$) of student $i$, in cohort $c$, at time $t$, using the following model:

$$\theta_{ict}^{EAP} = \beta_0 + \beta_1 \text{Grade}_{ict} + \beta_2 \text{SE}_{Index_i} + \beta_3 \text{Grade} \ast \text{SE}_{Index}_{ict} + \delta_c + \epsilon_{ict} \quad (3.2)$$

where $\text{Grade}_{ict}$ is the grade level of student $i$, in cohort $c$, at time $t$, $\text{SE}_{Index_i}$ is student $i$’s socioeconomic index, $\text{Grade} \ast \text{SE}_{Index}_{ict}$ is the interaction of the student’s grade level and his/hers SE index, $\delta_c$ is a cohort fixed effect and $\epsilon_{ict}$ is the stochastic error term. The main coefficient of interest is $\beta_3$, the interaction of grade level and the SE index. The regression results are presented in Table 3.7 in columns (3)-(4) and (7)-(8).

TABLE 3.7 HERE

The regression results suggest that high SE Index students on average have a higher ability score. In math, for the entire sample (column (3)) a one standard deviation increase in the SE Index is associated with a 0.04 increase in the ability score parameter and the coefficient is statistically significant at the 5% level. Furthermore, as students move through school, the gap between low- and high-SE Index students appears to be growing by 0.016 per grade. The initial gap of 0.04 in the ability score parameter between a student at the mean of the SE index and a student one standard deviation above the mean, will grow to a gap of 0.12 by grade 5. If we
exclude grade 1 from the analysis, the coefficient on the SE index increases to 0.058, but remains statistically significant at the 5% level. The coefficient on the interaction loses statistical significance, suggesting that most of the growth in the gap between low- and high-SE index takes place between grades 1 and 2 in math.

In contrast, there is no statistically significant gap in language between low- and high-SE index students as they enter school. However, with higher grade levels the gap grows by 0.026 per grade, per SE index score. Exclusion of grade 1 data shows that by the time students enter grade 2 there is a statistically significant ability score parameter gap of 0.063 between students one standard deviation of the SE index apart. Furthermore, the gap grows as the students enter higher grades, by 0.2 per grade level. The regression results show that there are significant inequalities in educational achievement between low- and high-SE index students and that those inequalities are exacerbated, as students enter higher grades.

3.5.3 Gaps

The analysis presented so far provides insight into how the students learn while at school, but it does not provide any information about how well the students do relative to a performance standard. We define the standard as the benchmark ability score which at a given grade level would result in a 50% test score in the entire universe of grade appropriate questions. We define the distance from the standard as the difference between the grade-appropriate standard and the student ability score parameter. By construction for students above the standard the distance will be negative. We model distance from the standard for student i in cohort c, at time t, in the ability score decile j:

\[
\text{Distance}_{ictj}^{EAP} = \beta_0 + \beta_1 \text{Grade}_{it} + \sum_j (\gamma_j \text{Grade}_i \times \text{Dec}_{ictj}^{EAP} + \alpha_j \text{Dec}_{ictj}^{EAP}) + \delta_c + \varepsilon_{ictj} \tag{3.3}
\]
where Grade$_{it}$ is student i’s grade level in year t, $Dec_{ictj}^{EAP}$ is a binary variable equal to one if student i in cohort c at time t is in decile j of the ability score distribution. We exclude deciles 5 and 6 from the estimation, so that the results can be interpreted as relative to the students in the middle of the distribution. The $Grade \times Dec_{ictj}^{EAP}$ variable is an interaction of the Grade and Dec variables discussed above, $\delta_c$ is a cohort fixed effect and $\varepsilon_{ictj}$ is the stochastic error term. We estimate the model using robust standard errors, clustered at the school level. The main coefficients of interest are the coefficients on Grade and the interactions of grade and decile ($\beta_1$ and $\gamma_j$). We further test whether the sum of the coefficients on grade and Grade*Dec are statistically different from zero. The regression results are presented in Table 3.8. Given the issues associated with the grade 1 gains discussed previously, we present the results for grades 2-5, but the results for the entire sample can be provided upon request.

TABLE 3.8 HERE

The analysis in math shows that by the time students enter grade 2 they are far below the standard, as the intercept is positive and statistically significant. Furthermore, for each grade level, the students in the middle of the distribution fall further behind the grade appropriate standard as the gap between the standard and the student ability score parameter increases by 0.176 points per grade. The regression results further show that the students in the bottom four deciles fall behind the standard even faster than the middle of the distribution. In contrast, the students at the top of the distribution fall behind the standard slower than the students in the middle. The students in the top decile fall behind by 0.061 points per grade level, as compared to the students in the middle falling behind at a rate of 0.176 points per grade level, and those in the bottom decile with a rate of 0.208 points per grade level.
In language the distance from the standard for the students in the middle of the distribution appears to be increasing only slightly as they progress to higher grades; the coefficient on grade level is 0.006, but is statistically significant at the 1% level. Similarly to math, the gaps between the standard and the student ability score parameters increase with grade level for the students in the bottom of the achievement distribution. However, for the student in the top four deciles, the distance from the grade-appropriate standard is falling, as students move to higher grades. The results suggest that in language, the top of the distribution is slowly catching up to the syllabus expectations.

3.5.4 Flattening/learning to read

The results presented so far highlighted the fact that the largest gains and the biggest increases in inequality between students in general and based on the SE Index take place between grades 1 and 2. Formally, this observation is solidified in Table 3.7. In all of the columns, if we compare the coefficient on the grade level with and without the grade 1 data, we can see that the estimated trajectories for grades 2 through 5 are more flat than those estimated with the inclusion of grade 1 (0.246 as compared to 0.092 for columns (1) and (2), and 0.269 as compared to 0.160 for columns (5) and (6)). In all cases, the coefficients are statistically significant. The results suggest that after the initial boost in performance, there is a significant slowing down in student learning.

As previously postulated, the main reason for the slowdown might be the fact that in higher grades students have to know how to read in order to answer the test questions, while in grade 1 and in part in grade 2, the test questions are read out loud to the students. The hypothesis is further reinforced by the observations made with regards to the learning trajectories, where we can observe a positive trajectory for
grade 1 material, but flat trajectories for higher grade level material. If the students do not learn to read when they are expected, it becomes very difficult to master any new material, as the students move through grades.

### 3.5.5 Productivity of the school system

So far we have shown that there appears to be very little learning taking place in relation to the material related to grades 2 through 5 in math, and that only students in the top of the ability distribution keep up with the syllabus in language. The results suggest that there is little benefit for attending additional years of schooling for a large fraction of the student population. In order to investigate this claim further we investigate how an additional year of schooling contributes to students ability to answer a given grade levels question correctly. If staying in school helps students answer questions from previous grades material, then they might not be learning according to the syllabus expectations, but they might still be making significant gains. Table 3.9 shows the percentage of students who could answer a grade N question in grade N+1, who could not answer the same question in grade N, as a share of students who could not answer the question in grade N.

Table 3.9 HERE

Similar to the rest of our analysis, the remedial effects of an additional year of schooling are concentrated in the grade 1 material, while there is very little gain in relation to grade 2 through grade 4 material. In math 27.4% of students in the 5-year cohort, who could not answer a grade 1 question, were able to answer a grade 1 question correctly by grade 2. In language it is 26.83% of students. The average gain related to grade 1 material is equal to 15.03%, however, if we exclude the gains
made between grades 1 and 2, the average falls to 10.91%. In language, the data shows a similar pattern - the average gain on grade 1 material is 12.64%, and with the exclusion of grade 2 gains, it falls to 7.91%. The results align with the previous observations about flat trajectories of learning on grade 2 through grade 4 material and further highlight the fact that the trajectories for grade 1 material flatten by grade 2. The results suggest that students gain very little after grade 2, putting the productivity of the rural public primary schools into serious question.

### 3.5.6 Progress Over Time

We have shown that when already in school, there is very little learning happening after grade 2. The longitudinal data combined with the vertically-scaled IRT model allows us to further investigate whether the system improves over time. To do so we model the ability score parameter of student $i$ in grade $g$ at time $t$ in the following way:

$$\theta_{igt}^{EAP} = \beta_0 + \beta_1 Year_t + \delta_g + \varepsilon_{igt} \quad (3.4)$$

where $Year_t$ is the year, $\delta_g$ is a grade fixed effect and $\varepsilon_{igt}$ is a stochastic error term. Controlling students grade level, the coefficient on the year can be interpreted as the overall effect of the changes to the schooling system. If the coefficient is positive, the regression would suggest that the system is improving over time. The regression results are presented in Table 3.10.

TABLE 3.10 HERE

Regression results suggest improvement in student ability scores over time; the coefficient on year is 0.07 and is statistically significant. The results might reflect improvements outside of the school; therefore we introduce students SE index
to capture those changes, at least in part. With the introduction of the SE index, there still appear to be improvements in student ability parameter scores in relation to math. However, it is possible that the estimated effects are associated with unobserved factors, not related to the school system at all. In language, there is no change over time in relation to student ability score parameters, whether or not we control for students SE index.

3.6 Conclusions

Low-income countries have significantly increased their education spending in the past couple of decades in an attempt to meet the MDG of achieving universal primary education. However, a growing body of evidence suggests that learning levels in many low-income countries are very low in spite of strong gains in school enrollment. It is therefore striking that there has been no systematic effort to date to document the extent to which students are learning during their time in school. We present the first estimates of learning trajectories in a low-income setting by following a representative sample of a cohort of rural public primary school students in the Indian state of Andhra Pradesh, through five years of primary school and show that vertically-scaled IRT models work well enough for such models to be used to estimate trajectories at the school-grade and district-grade level. Our results provide several insights into the performance of education systems in low-income settings.

First, we show that learning trajectories in low-income settings are considerably flatter than the rate of progress envisaged in the syllabus and that the gap between average student ability levels and the grade-appropriate standard grows over time (especially for math). Second, we find a significant increase in learning inequality within cohorts over time, and that this is correlated with socioeconomic status of households. Third, we find that the marginal product of an additional year in school
after the second grade is very low. In particular, the sharp flattening of learning profiles after second grade suggests that students who are not equipped with basic skills (especially reading) in early grades are unlikely to learn much from spending additional time in school.

Our results are consistent with recent hypotheses that education systems in low-income settings may be adversely affected by over-ambitious curricula that try to cover too much material and move on forward regardless of whether students have comprehended previous materials (Pritchett and Beatty (2012)). A similar hypothesis has been suggested by Banerjee and Duflo (2013) as a way of explaining why expensive expansions of inputs such as infrastructure and class-size reductions have not been very effective at improving test scores, while inexpensive programs of supplemental instruction delivered by community volunteers have been very effective at improving learning outcomes. They contend that the key mechanism for the effectiveness of the supplemental teaching interventions is that the teaching is targeted at the level of where the student is as opposed to where the textbook is. Our results are also consistent with recommendations made in recent reviews of the literature (see Muralidharan and Sundararaman (2013)) that it may be very important (and effective) to provide supplemental and customized instruction to in early grades to enable all children to catch up to the point of being able to read to learn by the time they enter grade 3.

More broadly, our results are consistent with the idea that curricula in these settings have been designed by highly educated elites and reflect a period of time when there was no expectation of universal primary education. Indeed, the historical purpose of education systems in many developing countries may not have been to provide 'human capital' to all students as much as to screen gifted students for positions of responsibility in the state and the clergy. Since the teachers continue
to follow the textbook as the default mode of instruction, and define their goals in terms of completing the curriculum over the course of year, it is not surprising that they are effectively 'teaching to the top' of the distribution and that a large number of children are in the class but not learning because the lesson is too advanced for them. Thus, our results suggest that the syllabi and pedagogy of education systems in these settings may not have effectively made the transition from serving a screening function to being able to add human capital at all parts of the student achievement distribution.

Documenting learning trajectories over time can be a critical source of feedback to policy makers on the state of education systems and on its effectiveness at different parts of the achievement distribution. We hope that the combination of the methods validated in this paper and the rich insights yielded by our longitudinal data will be followed up by similar efforts in other settings that can both shed light on the longitudinal performance of education systems and help direct supplemental resources where they are most needed.

3.7 Item Response Theory- Appendix 1

The technical appendix takes advantage of the IRT fundamentals as presented in Das and Zajonc (2010).

Item Response is a framework characterizing any test question using a set of up to three parameters to express the probability of a correct answer. The function using the three parameters is referred to as an Item Response Function. We focus our attention on a three parameter logistics model (3PL) as it allows to characterize response patterns for questions including multiple choice. Question q can be described using three parameters: a discrimination parameter \(a_q\), a difficulty parameter \(b_q\),
and a guessing parameter \( (c_q) \).

A guessing parameter \( (c_q) \) specifies the probability of a correct answer if the question is answered purely by guessing. A difficulty parameter \( (b_q) \) characterizes the student latent ability score required for the student to obtain a \( \frac{1+c_q}{2} \) probability of a correct answer. Finally, the discrimination parameter is the slope of the Item Response Function at the steepest point \( (b_q, \frac{1+c_q}{2}) \). The three parameters can be formalized using a logistics functional form to produce an Item Characteristics Curve (ICC), estimating a probability of a correct answer:

\[
ICC_q(\theta) = c_q + \frac{1-c_q}{1+e^{-D*a_q(\theta-b_q)}}
\]  

(A.3.1)

Graphically, an example of an ICC is presented in Figure A.3.1:

![Item Characteristics Curve](image)

**Figure A.3.1:** A Representative ICC Curve; \( a_q=1.25, b_q=0, c_q=0.25 \)

Item Response is a framework mapping student’s latent ability score \( (\theta) \), through the use of ICCs to a probability of a correct answer to any question with a set of parameters \( (a_q, b_q, c_q) \), as described in Equation A.3.1. Conventionally, it is assumed that the knowledge domain tested is one dimensional, i.e. student’s
latent ability score ($\theta$) is the only characteristic of the student that determines student’s responses. Furthermore, the framework assumes local independence, where the responses to two questions are not correlated with each other except through $\theta$. The IRT framework, under the two assumptions mentioned, allows to characterize a likelihood of a given pattern of responses ($S_i$) across N individuals as:

$$Pr(S = s|\theta) = \prod_{q=1}^{j} \prod_{i=1}^{N} ICC_q(s_{iq}|\theta_i)$$

$$= \prod_{q=1}^{j} \prod_{i=1}^{N} ICC_q(\theta_i)^{s_{iq}} [1 - ICC_q(\theta_i)]^{1-s_{iq}}, \text{ where } s_{iq} \in \{0, 1\}$$

where $s_{iq}$ is an indicator for whether the response provided by student i to question q was a correct one or not.

There are several approaches to estimating IRT models. We take advantage of the maximum likelihood estimation (MLE) and the expected a posteriori (EAP) approach. The MLE approach solves Equation A.3.2 as a two-step problem. First, item parameters are estimated, and later are used to estimate the student latent ability scores. The procedure integrates out the latent abilities and estimates the item parameters of the conditional maximum likelihood in the first step. Alternatively, EAP approach takes advantage of Bayesian-style updating, where item parameters are estimated and combined with a prior about the distribution of latent ability scores, and the pattern of responses, in order to obtain an improved, a posteriori measures of parameters of interest.

Given a set of parameters for any question in our data bank, we can use the estimated latent student ability scores in order to predict the likelihood of a correct answer to any question, even one that a particular student has never encountered.
The predicted probabilities allow us to track student progress in primary school as measured using the entire universe of grade-appropriate questions, providing a richer measure of student achievement, than one obtained only from questions the student actually answered.

### 3.8 Tables

**Table 3.1:** Distribution of Students Across Years and Grades in the Longitudinal Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1,518</td>
<td>1,614</td>
<td>1,325</td>
<td>1,358</td>
<td>1,315</td>
<td>7,130</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1,625</td>
<td>1,639</td>
<td>1,451</td>
<td>1,383</td>
<td>1,242</td>
<td>7,340</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1,676</td>
<td>1,649</td>
<td>1,591</td>
<td>1,493</td>
<td>1,344</td>
<td>7,753</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1,824</td>
<td>1,667</td>
<td>1,580</td>
<td>1,624</td>
<td>1,459</td>
<td>8,154</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1,949</td>
<td>1,909</td>
<td>1,621</td>
<td>1,633</td>
<td>1,614</td>
<td>8,726</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>8,592</td>
<td>8,478</td>
<td>7,568</td>
<td>7,491</td>
<td>6,974</td>
<td>39,103</td>
</tr>
</tbody>
</table>

**Table 3.2:** Special Assessment Test Design

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Grade</th>
<th>Common</th>
<th>Adjacent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>A 10</td>
<td>B 10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>A 10</td>
<td>B 10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>A 10</td>
<td>B 10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>A 10</td>
<td>B 10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>A 10</td>
<td>B 10</td>
<td>15</td>
</tr>
</tbody>
</table>
Table 3.3: Out-of-Sample Predictive Power of the IRT Model in the Cross Sectional Data

<table>
<thead>
<tr>
<th>Actual Score in Mathematics</th>
<th>Observation Level</th>
<th>Grade-Student</th>
<th>Grade-School</th>
<th>Grade-District</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Score</td>
<td></td>
<td>0.990**</td>
<td>0.965**</td>
<td>0.966**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.008)</td>
<td>(0.017)</td>
<td>(0.060)</td>
</tr>
<tr>
<td>Observations</td>
<td>3,940</td>
<td>291</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>R-squared</td>
<td>0.851</td>
<td>0.905</td>
<td>0.875</td>
<td></td>
</tr>
</tbody>
</table>

Language

| Predicted Score             |                   | 0.716**       | 0.801**      | 0.877**       |
|                             |                   | (0.015)       | (0.031)      | (0.091)       |
| Observations                | 3,943             | 291           | 15           |
| R-squared                   | 0.648             | 0.744         | 0.861        |

Robust standard errors in parentheses

** p < 0.01, * p < 0.05

Table 3.4: Within-Sample Predictive Power of the IRT Model in the Longitudinal Data

<table>
<thead>
<tr>
<th>Actual Score in Mathematics</th>
<th>Observation Level</th>
<th>Grade-Year</th>
<th>Grade-School-Year</th>
<th>Grade-District-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Score</td>
<td></td>
<td>1.054**</td>
<td>1.056**</td>
<td>1.036**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.003)</td>
<td>(0.006)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Observations</td>
<td>37,087</td>
<td>2,440</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>R-squared</td>
<td>0.897</td>
<td>0.964</td>
<td>0.982</td>
<td></td>
</tr>
</tbody>
</table>

Language

| Predicted Score             |                   | 1.006**    | 1.028**          | 1.036**             |
|                             |                   | (0.002)    | (0.004)          | (0.005)             |
| Observations                | 37,161            | 2,439      | 125               |
| R-squared                   | 0.918             | 0.971      | 0.995             |

Robust standard errors in parentheses

** p < 0.01, * p < 0.05
Table 3.5: Fraction of Students At or Above a 50% Score in the 5-Year Cohort

<table>
<thead>
<tr>
<th>Question Level</th>
<th>Student Grade Level</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>Grade 1</td>
<td>2.44%</td>
<td>36.24%</td>
<td>35.60%</td>
<td>46.58%</td>
<td>63.26%</td>
</tr>
<tr>
<td></td>
<td>Grade 2</td>
<td>16.05%</td>
<td>18.18%</td>
<td>23.94%</td>
<td>25.46%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade 3</td>
<td>10.94%</td>
<td>13.51%</td>
<td>16.11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade 4</td>
<td>5.86%</td>
<td>7.62%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.77%</td>
</tr>
<tr>
<td>Language</td>
<td>Grade 1</td>
<td>11.45%</td>
<td>46.46%</td>
<td>50.25%</td>
<td>58.95%</td>
<td>62.78%</td>
</tr>
<tr>
<td></td>
<td>Grade 2</td>
<td>28.02%</td>
<td>31.22%</td>
<td>38.21%</td>
<td>38.77%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade 3</td>
<td>31.47%</td>
<td>37.35%</td>
<td>37.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade 4</td>
<td>26.85%</td>
<td>29.53%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade 5</td>
<td></td>
<td>28.23%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The table presents the percentage of students reaching a minimum of 50% score on grade specific material at a given grade level.

Table 3.6: Pair-wise Leven’s Test for Equality of Variances of the Ability Scores in the 5-Year Cohort

<table>
<thead>
<tr>
<th>Math</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>0.455</td>
<td>0.402</td>
<td>0.638</td>
<td>0.704</td>
<td>0.722</td>
</tr>
<tr>
<td>Leven Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>33.81**</td>
<td>31.75**</td>
<td>72.25**</td>
<td>38.85**</td>
<td></td>
</tr>
<tr>
<td>Grade 2</td>
<td>11.58**</td>
<td>183.59**</td>
<td>121.59**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td></td>
<td>7.07**</td>
<td>0.932</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
<td>2.35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Telugu        |         |         |         |         |         |
| Variance      | 0.694   | 0.473   | 0.738   | 0.777   | 0.976   |
| Leven Test    |         |         |         |         |         |
| Grade 1       | 104.57**| 0.679   | 1.338   | 30.16** |
| Grade 2       | 75.79** | 119.46**| 217.74**|
| Grade 3       | 3.58    | 36.05** |
| Grade 4       |         | 18.43** |

Note: Main diagonal is a pair-wise Leven’s test for equality of variances of adjacent grades (e.g. (Gr.1, Gr.2)) other entries are pair-wise Leven’s tests for the two grades specified by the row and column combination; we report the test values with $H_0$: the variances of the two samples are equal.
Table 3.7: Evolution of Student Ability Scored Across Grade Levels in the Longitudinal Data

<table>
<thead>
<tr>
<th>EAP Ability Score</th>
<th>Mathematics</th>
<th></th>
<th></th>
<th></th>
<th>Language</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Gr.1</td>
<td>Gr.1</td>
<td>Gr.1</td>
<td>Gr.1</td>
<td>Gr.1</td>
<td>Gr.1</td>
<td>Gr.1</td>
<td>Gr.1</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
</tr>
<tr>
<td>Grade</td>
<td>0.246**</td>
<td>0.092**</td>
<td>0.239**</td>
<td>0.087**</td>
<td>0.269**</td>
<td>0.160**</td>
<td>0.259**</td>
<td>0.152**</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.013)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.013)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>SE Index</td>
<td>0.040*</td>
<td>0.058*</td>
<td>0.039</td>
<td>0.063*</td>
<td>0.026**</td>
<td>0.020*</td>
<td>0.026**</td>
<td>0.020*</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.024)</td>
<td>(0.021)</td>
<td>(0.028)</td>
<td>(0.007)</td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Grade*SE Index</td>
<td>0.016*</td>
<td>0.011</td>
<td>0.026**</td>
<td>0.020*</td>
<td>0.011</td>
<td>0.026**</td>
<td>0.020*</td>
<td>0.026**</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.886**</td>
<td>-0.289**</td>
<td>-0.883**</td>
<td>-0.288**</td>
<td>-0.871**</td>
<td>-0.419**</td>
<td>-0.864**</td>
<td>-0.414**</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(0.021)</td>
<td>(0.118)</td>
<td>(0.035)</td>
<td>(0.173)</td>
<td>(0.046)</td>
<td>(0.190)</td>
<td>(0.069)</td>
</tr>
<tr>
<td>Cohort FE</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Observations</td>
<td>26,770</td>
<td>21,725</td>
<td>26,770</td>
<td>21,725</td>
<td>26,834</td>
<td>21,757</td>
<td>26,834</td>
<td>21,757</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.107</td>
<td>0.021</td>
<td>0.118</td>
<td>0.035</td>
<td>0.173</td>
<td>0.046</td>
<td>0.190</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Robust standard errors, clustered at the school level

** p < 0.01, * p < 0.05
Table 3.8: Distance of Student’s EAP Ability Score from the Grade Level Standard in the Longitudinal Data in Grades 2-5

<table>
<thead>
<tr>
<th></th>
<th>Mathematics</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>0.176**</td>
<td>0.006**</td>
</tr>
<tr>
<td>Grade* Decile 1</td>
<td>0.032**</td>
<td>0.111**</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Grade* Decile 2</td>
<td>0.063**</td>
<td>0.090**</td>
</tr>
<tr>
<td></td>
<td>(0.004)††</td>
<td>(0.003)††</td>
</tr>
<tr>
<td>Grade* Decile 3</td>
<td>0.056**</td>
<td>0.055**</td>
</tr>
<tr>
<td></td>
<td>(0.004)††</td>
<td>(0.003)††</td>
</tr>
<tr>
<td>Grade* Decile 4</td>
<td>0.038**</td>
<td>0.029**</td>
</tr>
<tr>
<td></td>
<td>(0.003)††</td>
<td>(0.002)††</td>
</tr>
<tr>
<td>Grade* Decile 7</td>
<td>-0.037**</td>
<td>-0.021**</td>
</tr>
<tr>
<td></td>
<td>(0.002)††</td>
<td>(0.002)††</td>
</tr>
<tr>
<td>Grade* Decile 8</td>
<td>-0.064**</td>
<td>-0.040**</td>
</tr>
<tr>
<td></td>
<td>(0.002)††</td>
<td>(0.003)††</td>
</tr>
<tr>
<td>Grade* Decile 9</td>
<td>-0.091**</td>
<td>-0.056**</td>
</tr>
<tr>
<td></td>
<td>(0.003)††</td>
<td>(0.003)††</td>
</tr>
<tr>
<td>Grade* Decile 10</td>
<td>-0.115**</td>
<td>-0.066**</td>
</tr>
<tr>
<td></td>
<td>(0.007)††</td>
<td>(0.007)††</td>
</tr>
<tr>
<td>Constant</td>
<td>0.497**</td>
<td>0.287**</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.007)</td>
</tr>
</tbody>
</table>

|                | ✓ | ✓ |
| Decile FE      | ✓ | ✓ |
| Cohort FE      | ✓ | ✓ |
| Observations   | 31,942 | 31,947 |
| R-squared      | 0.957 | 0.953 |

Robust standard errors, clustered at the school level

Test for total effect Class+Class*Dec=0: † † p < 0.01, † p < 0.05

** p < 0.01, * p < 0.05
Table 3.9: Gains on Grade Appropriate Questions from Additional Years of Schooling in the Longitudinal Data in the 5-Year Cohort

<table>
<thead>
<tr>
<th>Level of the Question</th>
<th>Grade 2 Mean</th>
<th>Grade 3 Mean</th>
<th>Grade 4 Mean</th>
<th>Grade 5 Mean</th>
<th>Grade 3-5 Gain</th>
<th>Grade 4 Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>27.40%</td>
<td>9.88%</td>
<td>9.33%</td>
<td>13.51%</td>
<td>15.03%</td>
<td>10.91%</td>
</tr>
<tr>
<td>Grade 2</td>
<td>3.32%</td>
<td>2.66%</td>
<td>2.90%</td>
<td>2.96%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>2.02%</td>
<td>2.63%</td>
<td>2.33%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td></td>
<td>1.74%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>26.83%</td>
<td>8.58%</td>
<td>6.65%</td>
<td>8.50%</td>
<td>12.64%</td>
<td>7.91%</td>
</tr>
<tr>
<td>Grade 2</td>
<td>3.63%</td>
<td>2.76%</td>
<td>2.93%</td>
<td>3.12%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>2.62%</td>
<td>2.06%</td>
<td>2.34%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td></td>
<td>2.19%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The table presents the percentage of the total grade level population who are predicted to not answer the grade N question in grade N, but who could answer the question in grade N+1; e.g. 27.40% of students who could not answer a grade 1 math question in grade 1 could answer that question by the end of grade 2.

Table 3.10: Evolution of Student Ability Scored Across Years in the Longitudinal Data

<table>
<thead>
<tr>
<th>EAP Ability Score</th>
<th>Mathematics</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>0.070**</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>SE Index</td>
<td>0.090**</td>
<td>0.123**</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.297**</td>
<td>-0.007</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.049)</td>
</tr>
</tbody>
</table>

Grade FE ✓ ✓ ✓ ✓

Observations 26,770 26,770 26,834 26,834
R-squared 0.151 0.161 0.192 0.209

Robust standard errors, clustered at the school level
** p < 0.01, * p < 0.05
3.9 Figures

Figure 3.1: Graphical Representation of a Hypothetical Educational System

Figure 3.2: Comparison of Actual and Predicted Responses to Overlapping Common Questions in the Cross Sectional Data
Figure 3.3: Comparison of Actual and Predicted Responses in the Longitudinal Data

Figure 3.4: IRT Model Prediction Errors in the Longitudinal Data
Figure 3.5: Evolution of the EAP Ability Scores in the 5-Year Cohort

Figure 3.6: Predicted Probability of a Correct Answer on a Grade 1 Question in the 5-Year Cohort
Figure 3.7: Predicted Probability of a Correct Answer on a Grade 2 Question in the 4- and 5-Year Cohorts

Figure 3.8: Predicted Probability of a Correct Answer on a Grade 3 Question in the 3-, 4- and 5-Year Cohorts
Figure 3.9: Predicted Probability of a Correct Answer on a Grade 3 Question in the 2-, 3-, 4- and 5-Year Cohorts

Figure 3.10: Evolution of the EAP Ability Scores in the Longitudinal Data
3.10 Acknowledgements

I thank my co-author, Professor Karthik Muralidharan for his critical insight and expertise, invaluable in the process of writing this chapter. This paper is based on data collected for the Andhra Pradesh Randomized Evaluation Studies (AP RESt), which was set up as a research partnership between the Government of Andhra Pradesh, the Azim Premji Foundation, and the World Bank. I thank the AP RESt team (especially M Srinivasa Rao, J Umamaheswara Rao, B Srinivasulu, S Ramamurthy), and staff of the Azim Premji Foundation for collecting the data used in this paper, and thank Venkatesh Sundararaman for the overall support provided to the AP RESt project. Financial assistance for the data collection was provided by the Government of Andhra Pradesh, the UK Department for International Development (DFID), the Azim Premji Foundation, and the World Bank. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of any of the organizations that supported the data collection.

Chapter 3 is currently being prepared for submission for publication of the material. The dissertation author was the primary co-author of this material. Muralidharan, Karthik; Zieleniak, Jedrzej. “Chasing the Syllabus: Measuring Learning Trajectories in Developing Countries with Longitudinal Data and Item Response Theory.”
Chapter 4

Bibliography


Pritchett, L. and A. Beatty (2012): “The negative consequences of overambitious curricula in developing countries.”


